POLITÉCNICA

"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Graduado en Ingeniería Informática

Universidad Politécnica de Madrid

E.T.S de Ingenieros Informáticos

TRABAJO FIN DE GRADO

Modelos predictivos para el estudio del abandono en centros universitarios

AUTOR: Jesús Sánchez Santamaría

DIRECTOR: Victor Robles Forcada

MADRID, JUNIO DE 2014

Resumen

El abandono universitario es un problema que siempre ha afectado a las distintas universidades públicas. La Universidad Politécnica de Madrid, en concreto la Escuela Técnica Superior de Ingenieros Informáticos, pretende reducir dicho abandono aplicando técnicas de *Data Mining* para detectar aquellos alumnos que abandonan antes de que lo hagan. Este proyecto, a través de la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining) y con los datos obtenidos del SIIU (Sistema Integrado de Información Universitaria), tiene como fin identificar a los alumnos que abandonan el primer año para poder aplicar distintas políticas en ellos y así evitarlo.

Abstract

Early university abandonment is a problem which has affected different public universities. The Universidad Politécnica de Madrid (Polytechnic University of Madrid), in particular the Escuela Técnica Superior de Ingenieros Informáticos (Computer Science Engineering School), wants to reduce this abandonment rate using Data Mining techniques to find students who are likely to leave before they do. The purpose of this project, through CRISP-DM (CRoss Industry Standard Process for Data Mining) methodology and using the data from SIIU (Sistema Integrado de Información Universitaria), is to identify those students who leave the first year in order to apply different policies and avoid it.

"No tengas prisa"

Pablo Segóbriga

"Si haces lo que siempre has hecho,
obtendrás los resultados que siempre has obtenido."

Albert Einstein

Agradecimientos

A mis padres, por aguantarme y cuidarme tanto.
A mi hermano, por todas las veces que se tumba en mi cama.
A Alberto y Diego, por haber estado juntos hasta el final.
A Adrián, Jorge, Ignacio, Braulio y Andoni por seguir luchando en la FI.
A Mario, por todas las conversaciones que hemos tenido y aún nos quedan.
A Santiago, por sacarme de más de un apuro y por confiar a mí.
A Victor, por sus ganas de hacer cosas, aunque siempre llegue tarde.
A Ernestina, por todo lo que me ha enseñado.
A Sergio, por las partidas de ajedrez.
A Alex, por enseñarme a mirar las cosas desde otro ángulo.
A Laura, por estar siempre atenta.
A ACM (nueva y vieja guardia), por enseñarme a apasionarme con lo que hago.
A Jordi, Rebeca, Clara, Lucía, Geissell, Dani, Mireia, Laura y Karla, por cada momento (y los que nos quedan).
A Ana y a Diego, porque aunque no me entiendan, me escuchan.

A mis primos, por todo lo que he aprendo con vosotros.
A mis abuelos, por ser la voz que siempre habla desde la experiencia.
A mi familia, por estar siempre ahí.

A todos aquellos que han pasado por mi vida y que han dejado su granito de arena.

Índice general

Índice de tablas

Capítulo 1

Introducción

1.1. Introducción y motivación

Los últimos avances tecnológicos han logrado que los dispositivos de almacenamiento masivo (en relación precio – capacidad de almacenamiento), tales como los discos duros, puedan almacenar gigabytes de información a un precio reducido. Esto ha dado lugar a que las empresas y organismos almacenen todo tipo de información, desde los datos de sus clientes, sus transacciones o sus movimientos, hasta datos de telemetría, monitorización de pacientes o la evolución de los precios en los mercados.

En los comienzos, la ejecución de proyectos que involucraban *Data Mining* para el descubrimiento de conocimientos o proyectos, denominados *business inteligence*, fueron realizados mayoritariamente en el sector privado. Sin embargo, en los últimos años la situación se ha generalizado a todos los dominios y sectores, pasando a ser el sector público también el objetivo. De hecho, si se analizan los proyectos que en la actualidad se denominan de *Big Data*, destaca el sector público como uno de los retos.

La universidad pública como parte de este sector público almacena datos de todos sus estudiantes y el análisis de esta información permitiría poder establecer patrones que sirvieran, por ejemplo, para poder analizar el rendimiento que un alumno va a tener y así poder programar apoyos, poder establecer los horarios de manera coherente con los recursos y las necesidades, etc.

En concreto, lo que motiva el presente proyecto es la posibilidad de analizar datos de entrada de los alumnos a las titulaciones para poder extraer patrones que permitan predecir la probabilidad de abandono futuro de cara a poder enfocar recursos y actuaciones sobre colectivos de estudiantes de forma adecuada. Esto facilitará además una implementación adecuada de protocolos de calidad donde se incorporen actuaciones para reducir el indicador de abandono [García13].

1.2. Objetivos

Consecuentemente el objetivo de este trabajo es encontrar un modelo que sea capaz de describir y analizar el abandono de los alumnos de la Universidad Politécnica de Madrid (UPM). Este estudio permitirá aplicar un protocolo de prevención focalizado en un colectivo de estudiantes adecuado, con el fin de mejorar este indicador de calidad antes de que se produzca el abandono. Dicho protocolo consistirá en la implantación de distintas políticas tales como el Proyecto Mentor o las tutorías curriculares.

En concreto, se particularizará el análisis para el caso de aquellos que estén estudiando Grado de Ingeniería Informática en la Escuela Técnica Superior de Ingenieros Informáticos.

Se ha decido utilizar datos de los documentos oficiales enviados por la UPM al Sistema Integrado de Información Universitaria (SIIU). Los datos de cada cohorte de alumnos se obtienen de los archivos UPM de Matricula y Rendimiento de cada curso, junto con el archivo UPM de Matrícula del curso siguiente. Los datos de la cohorte 2009-2010, anterior al establecimiento del SIIU, se han tomado de Ágora UPM, con los mismos campos que los ficheros del SIIU.

Además, el fin de este proyecto es que se pueda repetir a lo largo de los años, por lo que otro de los objetivos es automatizar o, al menos, sistematizar la obtención de dichos modelos.

1.3. Estructura del documento

El resto del documento se ha estructurado de la siguiente manera:

- Estado de la cuestión. Se realizará un análisis tanto del proceso de KDD (*Knowledge Discovery in Databases*, Descubrimiento de información en Bases de Datos) como de la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining), sin olvidar tampoco los tipos de problemas de *Data Mining*.
- Comprensión del problema. En esta sección se hace la comprensión y el estudio del problema del abandono universitario comentado anteriormente.
- Modelado. En ella se evalúan los distintos modelos hallados.
- Sistematización. En este capítulo se explicará en detalle cómo poder repetir los resultados.
- Conclusiones y líneas futuras.

Capítulo 2

Estado de la cuestión

2.1. El proceso de KDD

El termino KDD (*Knowledge Discovery in Databases, Descubrimiento de información en Bases de Datos*), apareció por primera vez a finales de los años 80 [Piatesky91] para enfatizar que el conocimiento es el producto de un descubrimiento guiado por los datos, y ha servido como punto de unión para las diferentes áreas de investigación que se dedicaban a estudiar el proceso de análisis de datos y extracción de conocimiento a partir de los datos desde distintos puntos de vista, tales como el campo de las bases de datos, la estadística, las matemáticas, la lógica o la inteligencia artificial.

El proceso de KDD se define como *"El proceso no trivial de identificación de patrones válidos, nuevos, potencialmente útiles y comprensibles en los datos"* [Fayyad 96].

En esta definición los datos hacen referencia a un conjunto de hechos (casos de una base de datos). Un patrón hace referencia a una expresión en algún lenguaje que sirve para describir un subconjunto de los datos o a un modelo aplicable a esos datos. Es decir, un patrón es una instancia de un determinado modelo. Por ello, se entiende la extracción de patrones como la extracción de un modelo para unos datos, esto es, cualquier descripción de alto nivel de los datos.

El término proceso implica que KDD es la conjunción de muchos pasos repetidos en múltiples iteraciones. Se dice por otra parte que es no trivial, porque se supone que hay que realizar algún tipo de proceso complejo. Los patrones deben ser válidos, con algún grado de certidumbre, y novedosos, por lo menos para el sistema y, preferiblemente, para el usuario, al que deberán reportar alguna clase de beneficio (útil). Por último, está claro que los patrones deben ser comprensibles, si no de manera inmediata, sí después de alguna clase de preprocesado.

Todo lo indicado con anterioridad implica que se pueden definir medidas cuantitativas para evaluar los patrones obtenidos. Estas medidas pueden servir, por ejemplo, para evaluar la bondad, utilidad, simplicidad y certidumbre de los patrones. Las medidas de interés se pueden definir para ordenar los patrones obtenidos por un cierto sistema de

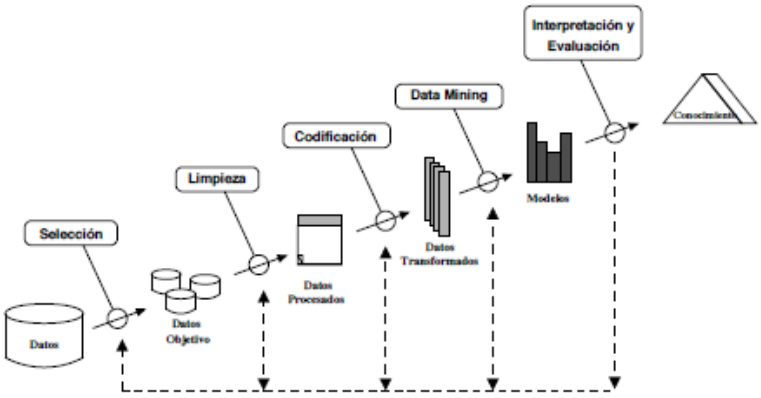

Figura 2.1: Fases del proceso de KDD [Frawley96]

KDD.

El proceso de KDD, según lo expuesto anteriormente, es el proceso de aplicar a una determinada base de datos, las operaciones requeridas de selección, preprocesado, muestreo, transformación y métodos de *Data Mining* para extraer los patrones y posteriormente evaluarlos para identificar el conjunto de ellos que representarán al conocimiento. La fase de *Data Mining* del proceso de KDD tiene que ver tan sólo con la aplicación de los algoritmos de extracción de patrones. El proceso total incluye, tanto la preparación previa, como las posibles fases de interpretación de los patrones obtenidos.

Data Mining es una etapa en el proceso de KDD que consiste en la aplicación de algoritmos de descubrimiento y de análisis de datos que bajo unas ciertas limitaciones de eficiencia computacional, produce una serie de patrones de los datos sobre los que actuó. Es importante resaltar que el espacio de posibles patrones es a menudo infinito, y que la enumeración de los mismos incluye alguna forma de búsqueda en ese espacio. Las restricciones computacionales en la práctica imponen límites severos en el subespacio de búsqueda a ser explorado por un cierto algoritmo de *Data Mining*.

El proceso de KDD es un proceso iterativo e interactivo porque incluye numerosos pasos en los que el usuario tiene que tomar decisiones. El proceso es iterativo porque puede ser necesario acceder desde una fase a cualquiera de las anteriores, e interactivo porque el proceso es supervisado y controlado por el usuario de forma directa.

El proceso de KDD consta de varias fases que van desde la selección de los objetivos hasta la interpretación y evaluación de los resultados obtenidos. En la Ilustración 1 pueden observarse las fases las que componen el proceso de KDD [Cabena 97]. A continuación se describen las fases del proceso de KDD.

- **Establecer el objetivo**

  En esta fase hay que estudiar el problema y decidir cuál es la meta del proyecto. Esta fase es similar a la de cualquier otro proyecto, y suele requerir la colabora-

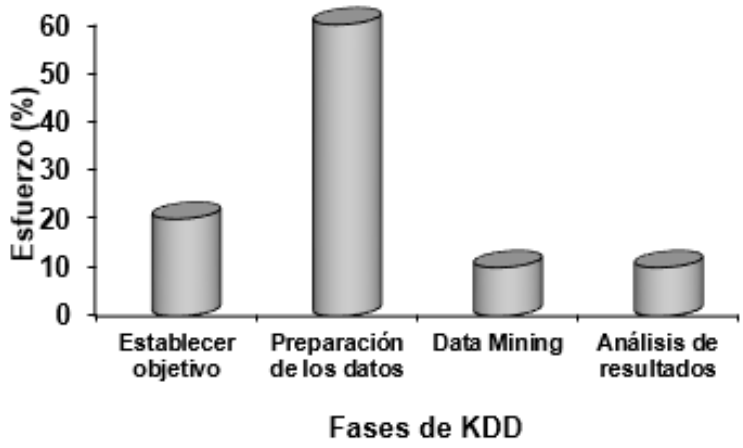

Figura 2.2: Estimación del esfuerzo dedicado a las fases de KDD [Frawley96]

ción de un experto. Si se hace bien el planteamiento del problema, se descubren fácilmente las fuentes de datos y los algoritmos de *Data Mining* que se aplicarán. Un mal planteamiento del problema nos puede llevar a resultados erróneos.

- **Preparación de los datos**

  Esta etapa del proceso de KDD es la que mayor esfuerzo requiere como se observa en la figura 2.2.

  En esta fase se pueden distinguir dos subfases fundamentales.

  1. **Compresión de los datos** Se identifican las fuentes de datos internas o externas y se selecciona el subconjunto de datos necesarios para la aplicación de *Data Mining*, ya sean tablas de una base de datos o ficheros de texto.
  2. **Compresión de los datos** Una vez identificados los datos a utilizar hay que estudiarlos para, por un lado entender el significado de los atributos y, por otro lado, para detectar errores de integración, como puede ser el hecho de que haya datos repetidos con distinto nombre o datos que significan lo mismo en diferente formato. Estos problemas surgen porque los datos pueden provenir de fuentes diferentes, y no todas almacenan la misma información de la misma manera. Con el preproceso de los datos lo que se consigue es tener un conjunto de datos adecuado para el correcto funcionamiento de las fases posteriores del proceso de KDD.

- **Data Mining**

  En esta fase se selecciona el algoritmo de *Data Mining* que vamos a aplicar así como los parámetros del algoritmo. Como cada algoritmo requiere un formato dife-

rente en los datos de entrada, en esta fase debemos transformar los datos para que se ajusten al formato de entrada del algoritmo seleccionado. Es la etapa donde se aplican los diferentes algoritmos de análisis de datos sobre los datos transformados y preparados en las etapas anteriores. Durante la etapa de *Data Mining* se buscan los patrones presentes en los datos. En función del algoritmo seleccionado se obtendrá un formato diferente en la salida. En esta fase es posible aplicar varias veces el mismo algoritmo o utilizar conjuntamente varios algoritmos.

- **Análisis de los resultados**

  En esta etapa es cuando se interpretan y evalúan los resultados obtenidos en la etapa de *Data Mining*. Se suelen utilizar técnicas de visualización para ver los resultados obtenidos. Una vez presentados los resultados el usuario debe interpretar los resultados y si no está de acuerdo debe volver a aplicar los algoritmos con otros parámetros, e incluso ejecutar otro algoritmo de *Data Mining*, para obtener unos resultados más reales. Todo esto hace que el proceso de KDD sea iterativo. En esta fase también se debe determinar cómo utilizar los resultados obtenidos. Los resultados se pueden integrar en un sistema experto o como procedimientos almacenados en un gestor de base de datos.

Es importante mencionar una vez más que si bien en los entornos académicos al proceso de descubrimiento se lo denominó KDD y *Data Mining* se dejó para nombrar la fase del proceso en la que los patrones de comportamientos son extraídos, hoy en día, *Data Mining* ha sido aceptado como término para referirse al proceso global si bien siempre hay que recordar que tal proceso en un proceso complejo compuesto de varias fases. En lo que sigue se analiza el primer estándar de proceso de *Data Mining* que ha aparecido el estándar CRISP-DM.

2.2. *Data Mining*: tipos de problemas.

2.2.1. Tipos de problemas de *Data Mining*

Tal y como se ha comentado con anterioridad existen innumerables problemas de *Data Mining* con los que nos podemos encontrar:

- Predecir el nivel de morosidad de un cliente
- Saber quiénes son mis clientes
- Encontrar los productos que más frecuentemente se compran juntos
- Encontrar el perfil del comprador del producto A

- Detectar los clientes que están cometiendo acciones fraudulentas
- Encontrar el perfil del cliente que me abandonará el mes siguiente
- Encontrar los síntomas de enfermedades que más a menudo aparecen juntos
- Analizar los canales de televisión que a menudo se contratan juntas
- Calcular el valor potencial de un cliente
- ...

Es fácil comprobar ante esta lista que no pretende ser exhaustiva sino sólo mostrar algunos ejemplos de problemas *Data Mining* frecuentes, sin embargo podemos establecer una primera clasificación de estos problemas en:

- ***Problemas Descriptivos***
- ***Problemas Predictivos***

2.2.1.1. Problemas descriptivos

Entendemos en este contexto como un problema descriptivo aquellos cuya meta es simplemente encontrar una descripción de los datos de estudio. Pertenecen a este tipo de problemas el ejemplo de conocer cuáles son los clientes de una organización (características de los mismos), o el encontrar los productos que frecuentemente se compran juntos o síntomas de enfermedades que se presentan juntos.

La meta de todos estos problemas es una descripción del conjunto de datos origen. No obstante, analizando estos ejemplos más en detalle observamos que si bien ambos pretenden descubrir características del conjunto origen en el primer caso (descripción de los clientes), lo que se pretende de alguna manera es agrupar a los clientes en grupos más o menos homogéneos y extraer las características de estos objetos. Sin embargo, en el segundo tipo de consultas (productos que se compran juntos o síntomas de enfermedades que se presentan juntos), si bien el problema sigue siendo descriptivo, el tipo de descripción requerida es diferente pues lo que se pretende es encontrar asociaciones esta vez no entre los objetos origen sino entre los valores de atributos o propiedades de estos objetos. Esto provoca una división más detallada del problema descriptivo en:

- ***Análisis de segmentacion***: Se refiere a los problemas donde la meta es encontrar grupos homogéneos en la población de objetos origen. A estos problemas se los denomina también problemas de aprendizaje no supervisado. El típico ejemplo de segmentación es realizar una segmentación de los clientes.
- ***Análisis de asociaciones***: Hace referencia a los problemas en los que se persigue obtener relaciones entre los valores de atributos de una base de datos. El ejemplo más típico es el de análisis de la cesta de la compra.

2.2.1.2. Problemas Predictivos

Por otra parte, existen consultas de *Data Mining* cuya meta es obtener un modelo que en un futuro pueda ser aplicado para predecir comportamientos. Este tipo de problemas es el que denominamos problemas predictivos o en entornos de inteligencia Artificial se denominan problemas de aprendizaje supervisado.

Es importante destacar aquí que el proceso de *Data Mining* es siempre un proceso inductivo en que no se realiza ninguna predicción de los datos. Se llaman problemas predictivos porque si bien las técnicas aplicadas para la obtención del modelo son técnicas de inducción sobre los datos de origen, el resultado (modelo) será aplicado para predecir. Esta distinción conviene tenerla clara.

No obstante una vez más podemos analizar estos problemas con más detenimiento para observar que, por ejemplo, en el primer caso la variable a predecir es una variable categórica (si compra o no un producto) y sin embargo, en el caso del préstamo, la variable a predecir es la probabilidad de devolución del mismo, es decir, una variable numérica.

Esta distinción en el tipo de variables que el modelo predecirá nos lleva a una distinción dentro de los problemas predictivos que es:

- ***Problemas de clasificación***: Hacen referencia a los problemas en los que la variable a predecir tiene un número finito de valores, esto es la variable es categórica. Un ejemplo de este tipo de problemas sería encontrar un modelo que a la vista de un histórico de clientes clasificados como "buenos", "regulares" y "malos", establezca qué tipo de cliente es uno nuevo.

- ***Problemas de predicción de valores***: Se refieren a los problemas en los que la variable a predecir es numérica. Por ejemplo, podríamos tener el caso de encontrar un modelo que establezca la probabilidad de que un cliente que está pidiendo un préstamo lo devolverá o no.

El hecho de hacer esta distinción es importante, pues dependiendo del tipo de problema así será la técnica que se utilizará para solucionarlo.

2.3. CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) [Wirth00] es una metodología para proyectos de minería de datos. Está descrita en términos de procesos jerárquicos, comprendiendo cuatro niveles de abstracción (de más genérico a más específico): *fases, tareas generales, tareas específicas* e *instancias de procesos*.

La metodología CRISP-DM distingue entre el *Modelo de Referencia* y la *Guía de Usuario*, siendo el Modelo de Referencia en donde se describe qué hacer en un proyecto de minería de datos y la Guía de Usuario cómo hacer dicho proyecto

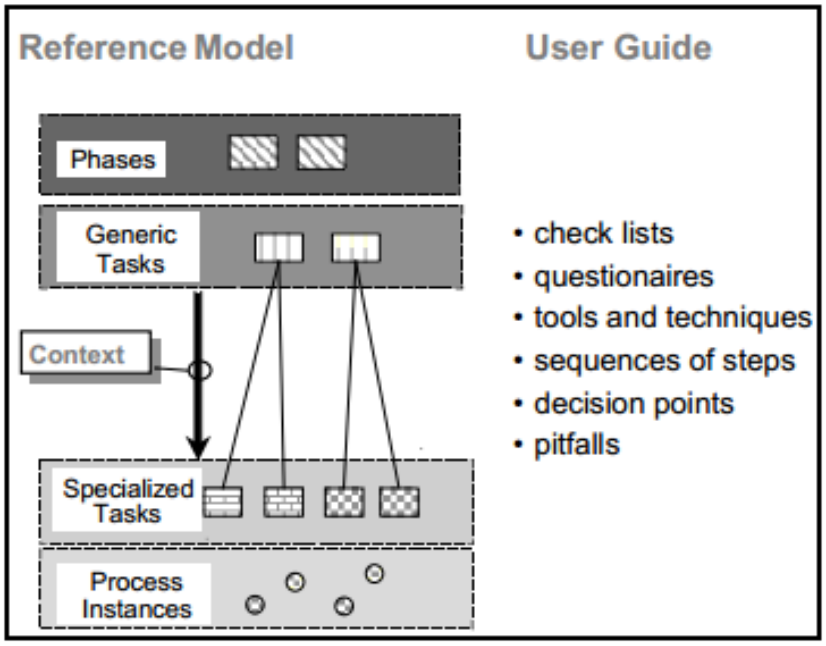

Figura 2.3: REFS: Los cuatro niveles de la metodología CRISP-DM (www.crisp-dm.org)

2.3.1. Modelo de Referencia Genérico de CRISP-DM

El modelo de referencia de CRISP-DM proporciona un resumen del ciclo de vida de un proyecto de minería de datos. Está dividido en seis fases, como muestra la Ilustración 4. La secuencia de las fases no sigue un orden estricto. Las flechas de la figura solo muestran las dependencias más frecuentes entre las fases. El círculo exterior simboliza la naturaleza cíclica de la minería de datos. Un proyecto de minería de datos no se acaba cuando se encuentra una solución; los modelos deben aprender de sí mismos y adquirir los conocimientos de sus antecesores.

Las fases de CRISP-DM son las siguientes:

- **Comprensión del problema**

  En esta fase inicial se realiza la comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, convirtiéndolo en una definición de un problema de minería de datos. Las tareas a realizar son las siguientes:

  - Determinar los objetivos del negocio.
  - Evaluar la situación.
  - Determinar las metas de *Data Mining*.
  - Producir el plan de proyecto.

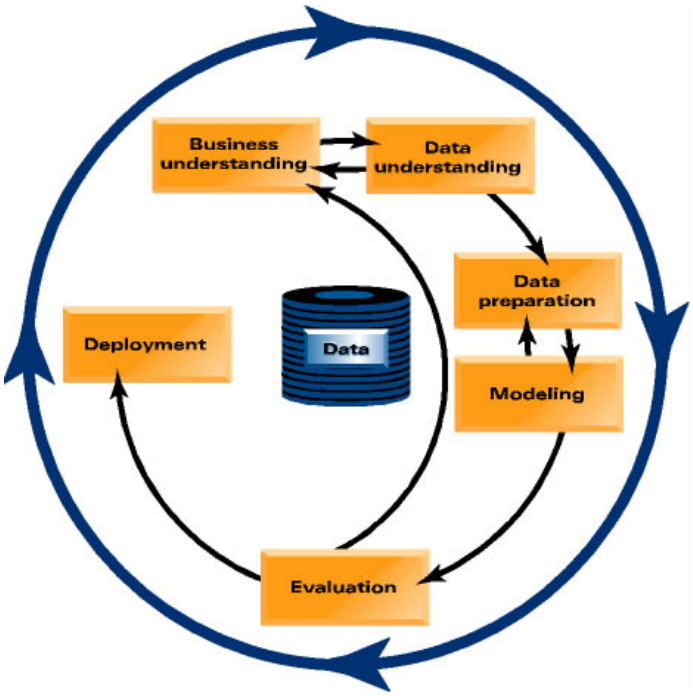

Figura 2.4: Fases de CRISP-DM (www.crisp-dm.org)

- **Comprensión de los datos**

  La comprensión de los datos comienza con una colección de datos inicial y realiza una familiarización con éstos, intentando de identificar los problemas de calidad con el fin de descubrir las primeras características de los datos o detectar subconjuntos para realizar las primeras hipótesis sobre la información oculta. La compresión de los datos tiene las siguientes tareas:

  - Conseguir el conjunto inicial de datos.
  - Estudio de los datos.
  - Establecimiento de los metadatos.
  - Establecimiento del tipo de variables (cuantitativas o cualitativas).
  - Establecer la caducidad de cada dato: vida de las variables.

- **Preparación de los datos**

  Esta fase cubre todas las actividades de construcción del conjunto final de datos (que serán los datos de entrada de los algoritmos de minería de datos) desde el conjunto inicial de datos. Es posible que estas tareas se tengan que realizar múltiples veces y sin orden determinado. Dichas tareas son las siguientes:

- Selección de datos
- Limpieza de datos
- Construcción del conjunto de datos
- Integración los datos
- Formato de los datos
- Transformación de los datos

- **Modelado**

  En esta fase se seleccionan distintas técnicas de minería y se aplican calibrando sus parámetros para conseguir los valores óptimos. Existen distintas técnicas para resolver el mismo problema, siendo la diferencia fundamental los requisitos que se han de cumplir en los datos de entrada, por lo que a menudo es necesario volver a la fase de preparación de los datos. El modelado cuenta con las siguientes tareas:

  - Selección de la técnica de modelado.
  - Generar un diseño de prueba.
  - Construir el modelo.
  - Evaluar el modelo.

- **Evaluación**

  En este fase se dispone de al menos un modelo que parece tener una buena calidad desde el punto de vista del análisis de datos. En este punto es importante cerciorarse de que también se han logrado los objetivos de negocio y si algún aspecto de éstos no se ha tenido suficientemente en cuenta. Al final de esta fase se tendrá la decisión sobre el uso de los resultados de minería. Las tareas a realizar son las siguientes:

  - Evaluar los resultados.
  - Proceso de revisión.
  - Determinar los pasos siguientes.

- **Implantación**

  La creación del modelo no es el final del proyecto. Éste se tiene que poner en orden y presentarlo de manera que se pueda hacer uso del mismo. Por tanto, esta fase puede ser tan simple como la generación de un informe o tan compleja como la implantación de un proceso de minería en toda la empresa. Las tareas que hay que llevar a cabo son las siguientes:

- Desarrollo del plan de implantación.
- Desarrollo del plan de monitorización y mantenimiento.
- Realización del informe final.
- Revisión del proyecto.

2.4. Herramientas utilizadas

2.4.1. Clementine

Clementine es un *software* de *IBM* usado para crear modelos de *Data Mining* de manera visual. Con la adquisición de *IBM* su nombre cambió a *IBM SPSS Modeler*.

2.4.2. Knime

Knime es un *software* libre bajo licencia GNU desarrollado por el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, sobre la plataforma *Eclipse* y programado en *Java*. Es usado para crear modelos de *Data Mining* de manera visual, con una serie de nodos que se despliegan e interactúan de manera gráfica.

2.4.3. Weka

Weka es un *software* libre bajo licencia GNU desarrollado en la Universidad de Waikato, Nueva Zelanda. Está escrito en *Java* y sirve para el aprendizaje automático y *Data Mining*. Se ha utilizado como un *plugin* de Knime, pudiendo utilizar sus funciones como nodos.

2.4.4. Microsoft Excel

Microsoft Excel es un *software* de *Microsoft Windows* usado para crear hojas de cálculo que forma parte de las aplicaciones de *Microsoft Office*.

2.4.5. Xampp

XAMPP es un paquete de *software* libre bajo licencia GNU, compuesto por el servidor APACHE HTTP y la base de datos MySQL, además de otras herramientas. Es un servidor web fácil de usar y gratuito, pudiendo ser utilizado en los sistemas operativos *Windows*, *Linux*, *Sun Solaris* y *Mac OS X*.

2.4.6. Google Docs

Google Docs & Spreadsheets es un programa gratuito basado en Web para crear documentos en línea. Incluye un procesador de textos, una Hoja de cálculo, un programa de presentación básico, un creador de dibujos y un editor de formularios destinados a encuestas. Además, *Google* también ofrece un servicio de alojamiento de archivos con *Google Docs* en el que guardar dichos documentos.

2.4.7. LaTeX

LaTeXes un procesador de textos que está formado mayoritariamente por órdenes construidas a partir de mandatos de TeX. Permite a quien escribe centrarse exclusivamente en el contenido, sin tener que preocuparse de los detalles del formato. Además, la salida que ofrece es siempre la misma, con independencia del dispositivo o el sistema y puede ser exportado muy fácilmente a partir de una misma fuente a numerosos formatos como *Postscript*, PDF, SGML, HTML o RTF.

Capítulo 3

Compresión del problema.

3.1. Definición

En este punto se establecerán los distintos objetivos del problema, así como los distintos resultados obtenidos del mismo, basándose en la metodología CRISP-DM.

3.2. Determinar los objetivos de negocio.

- **Objetivos del negocio**: Describir el abandono de alumnos en la universidad y conocer las variables y condiciones que hacen que sea más probable que un alumno abandone al comenzar sus estudios con el fin de evitar dicho abandono con distintos métodos o proyectos que involucren al alumno.
- **Criterios de éxito del proyecto**: Con la información obtenida del SIIU (Sistema de Integrado de Información Universitaria) se desea obtener un modelo que pueda predecir el abandono del estudiante al entrar en la universidad. El modelo debe maximizar el porcentaje de "bien clasificados" y minimizar aquellos que predice como "no abandono" pero finalmente abandonan.

3.3. Evaluar la situación.

Para la resolución del problema se dispone de las siguientes bases de datos:

- Tablas de Acceso a Grado del año 2009-2010
- Tablas de Acceso a Grado del año 2010-2011
- Tablas de Acceso a Grado del año 2011-2012

- Tablas de Acceso a Grado del año 2012-2013
- Tablas de Avance de alumnos de Grado del año 2013-2014
- Tablas de Rendimiento de alumnos de Grado del año 2009-2010
- Tablas de Rendimiento de alumnos de Grado del año 2010-2011
- Tablas de Rendimiento de alumnos de Grado del año 2011-2012
- Tablas de Rendimiento de alumnos de Grado del año 2012-2013

3.4. Determinar las metas de *Data Mining*

- **Metas de *Data Mining***
  - Encontrar un modelo que sea capaz de predecir el abandono de los alumnos del Grado en Ingeniería Informática en el primer año a partir de las variables Familia Numerosa, Niveles de estudio Padre, Niveles de estudio Madre, Ocupación Padre, Ocupación Madre, Forma de Admisión, Estudio Acceso, Naturaleza Centro Secundaria, Nota de Admisión, Rezagado, Centro Secundaria Madrid, Trabaja Estudiante, Sexo, Español y Edad.
- **Criterios de éxito**: Se considerará un modelo bueno aquel que llegue al 80 % (± 2 %) de bien clasificados, siempre y cuando los clasificados como *no abandono* que finalmente abandonan no superen el 5 %.

Objetivo de negocio	Objetivo de *Data Mining*	Criterios de éxito
Conocer la probabilidad de abandono futuro en el Grado en Ingeniería Informática nada más entrar en la Universidad	Tener un modelo que sea capaz de predecir el abandono del Grado en Ingeniería Informática nada más entrar en la Universidad.	Se considerará un modelo bueno aquel que llegue al 80 % (±2 %) de bien clasificados, siempre y cuando los clasificados como *no abandono* que finalmente abandonan no superen el 5 %.

3.5. Comprensión de los datos

3.5.1. Conjunto inicial de datos

Se han analizado datos de todos los alumnos de grado de los años 2009-2010, 2010-2011 y 2011-2012 en distintos momentos del curso, incluyendo datos personales y

académicos. (Ficheros entregados por la UPM al SIIU, excepto curso 2009-2010, que se han consultado los mismos campos). Los datos, según el SIIU, están distribuidos en distintas tablas:

- ***Tabla de Acceso***: En esta tabla aparece información de estudiantes que se matriculan por primera vez en grado. Se incluyen también aquellos estudiantes procedentes de otro grado así como los que se trasladen expedientes de plantes antiguos. En la tabla del año 09-10 hay 2431 alumnos, en la del año 10-11 hay 7036 y en la del 11-12 hay 7067. El año 09-10 hay menos porque fue el primer curso de implantación de los grados, y no se lanzaron todos.
- ***Tabla de Avance***: Incluye características generales de los estudiantes. Toda la información que aparece en las tablas de avance es información que se puede encontrar en las tablas de rendimiento y acceso, por lo que no se ha trabajado con ellas.
- ***Tabla de Rendimiento***: Aparece información del rendimiento académico de todos los estudiantes matriculados en el curso en estudios de grado impartidos en centros propios de la universidad. En la tabla del año 09-10 hay 2133 alumnos, en la del año 10-11 hay 8749 y en la del 11-12 hay 14884.

Cabe destacar que también disponemos de la tabla de avance del año 2012-2013.
A continuación se muestra una vista en detalle de cada tabla (SIIU):

- ***Tabla de Acceso***. Incluye las siguientes variables:
  - **Universidad.**Código de la universidad en la que se imparte el Grado.
  - **Centro**.Código del centro en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Municipio**.Código del municipio en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Campus**. Código del campus en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Titulación**. Código del Grado en el que está matriculado el estudiante.
  - **Tipo de documento identificativo** [**TipoDocIdentidad**]. Tipo de documento que presenta el estudiante. Valores:
    - 1 = NIF
    - 2 = NIE
    - 3 = Pasaporte
    - 4 = Otros

- **Nº de documento identificativo [NumeroDocumento]**. Número del documento que presenta el estudiante. El formato será:
  - NIF = nnnnnnnnL(donde nnnnnnn serán los 8 número y L la letra final que corresponda)
  - NIE = AnnnnnnnL (A + 7 números + letra final que corresponda)
  - Pasaporte u otros. El formato del pasaporte no se validará.
- **Año de nacimiento [AnioNacimiento]**. Los cuatro dígitos del año de nacimiento
- **Nacionalidad [PaisNacionalidad].** Código del país de nacionalidad.
- **Sexo.**
  - H = Hombre
  - M = Mujer
- **Modalidad reconocida de dedicación al estudio [DedicacionEstudio]**
  - 1 = Modalidad tiempo completo
  - 2 = Modalidad tiempo parcial
  - 3 = Otros
- **Trabajo del estudiante [OcupacionEstudiante].** Categorías:
  - 00 = Ocupaciones militares
  - 01 = Directores/as y gerentes
  - 02 = Técnicos/as y profesionales científicos/as e intelectuales
  - 03 = Técnicas/os y profesionales de apoyo
  - 04 = Empleadas y empleados de tipo contable y administrativo
  - 05 = Trabajadores y trabajadoras de los servicios de restauración, personales, protección y vendedores/as de los comercios
  - 06 = Trabajadores cualificados en la agricultura y en la pesca
  - 07 = Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria.
  - 08 = Operadores de instalaciones y maquinaria así como montadoras y montadores
  - 09 = Trabajadores no cualificados y trabajadoras no cualificadas
  - 10 = Parado o parada
  - 11 = Jubilados o jubiladas
  - 12 = Amas/os de casa
  - 13 = Incapacitados o incapacitadas para trabajar

  - 14 = Otra situación (Rentistas,...)
  - 88 = No aplica
  - 99 = No consta
- **Familia numerosa [FamiliaNum].**
  - 1 = No es familia Numero
  - 2 = Familia Numerosa General
  - 3 = Familia Numerosa Especial
- **Nivel de estudios del padre [NivelEstudioPadre].**
  - 1 = Analfabeto
  - 2 = Sin Estudios
  - 3 = Estudios Primarios
  - 4 = Estudios Secundarios
  - 5 = Estudios Superiores
  - 8 = No aplica
  - 9 = No consta
- **Nivel de estudios de la madre [NivelEstudioMadre].**
  - 1 = Analfabeto
  - 2 = Sin Estudios
  - 3 = Estudios Primarios
  - 4 = Estudios Secundarios
  - 5 = Estudios Superiores
  - 8 = No aplica
  - 9 = No consta
- **Trabajo del padre [OcupacionPadre].**
  - 00 = Ocupaciones militares
  - 01= Directores y gerentes
  - 02 = Técnicos y profesionales científicos e intelectuales
  - 03 = Técnicos y profesionales de apoyo
  - 04 = Empleados de tipo contable y administrativo
  - 05 = Trabajadores de los servicios de restauración, personales, protección y vendedores de los comercios
  - 06 = Trabajadores cualificados en la agricultura y en la pesca
  - 07 = Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria.

- 08 = Operadores de instalaciones y maquinaria y montadores
- 09 = Trabajadores no cualificados
- 10 = Parado
- 11 = Jubilado
- 12 = Amos de casa
- 13 = Incapacitados para trabajar
- 14 = Otra situación (Rentistas,...)
- 88 = No aplica
- 99 = No consta

- **Trabajo de la madre [OcupacionMadre].**
  - 00 = Ocupaciones militares
  - 01= Directores y gerentes
  - 02 = Técnicos y profesionales científicos e intelectuales
  - 03 = Técnicos y profesionales de apoyo
  - 04 = Empleados de tipo contable y administrativo
  - 05 = Trabajadores de los servicios de restauración, personales, protección y vendedores de los comercios
  - 06 = Trabajadores cualificados en la agricultura y en la pesca
  - 07 = Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción, y la minería, excepto los operadores de instalaciones y maquinaria.
  - 08 = Operadores de instalaciones y maquinaria y montadores
  - 09 = Trabajadores no cualificados
  - 10 = Parado
  - 11 = Jubilado
  - 12 = Amos de casa
  - 13 = Incapacitados para trabajar
  - 14 = Otra situación (Rentistas,...)
  - 88 = No aplica
  - 99 = No consta

- **Forma de admisión a este Grado universitarios [FormaAdmision].** Forma en la que el estudiante ha sido admitido en este grado según las siguientes categorías:
  - 01 = Pruebas de Acceso a la Universidad (PAU y Pruebas de Acceso anteriores)

- 02 =Estudiantes procedentes de sistemas miembros de la Unión Europea o de otros Estados con los que España haya suscrito Acuerdos Internacionales a este respecto que cumplan los requisitos exigidos en su respectivo país para el acceso a la universidad.
- 03 = Mediante posesión de los títulos de Técnica/o Superior correspondientes a las enseñanzas de Formación Profesional y Enseñanzas Artísticas o de Técnica/o Deportiva/o Superior correspondientes a las enseñanzas Deportivas o títulos equivalentes
- 04 = Mayores de 25 años
- 05 = Mayores de 40 años con acreditación de experiencia laboral o profesional
- 06 = Mayores de 45 años.
- 07 = Estudios procedentes de sistemas educativos extranjeros, previa solicitud de homologación, del título de origen al título español de bachiller
- 08 = Por poseer otro título universitario o título equivalente
- 09 = Convalidación parcial de estudios extranjeros (al menos 30 créditos reconocidos)
- 10 = Mediante traslado de Expediente proveniente de otro estudio de grado (al menos 30 créditos reconocidos)
- 11 = Incorporación desde enseñanzas anteriores a las establecidas por el RD 1393/2007.

- **Estudio de acceso a este Grado universitario [EstudioAcceso].**
  - Si se trata de un estudiante que accede por primera vez al SUE con forma de acceso 01 o 03.
    - 01 = Bachillerato LOE
    - 02 = Bachillerato LOGSE
    - 03 = COU
    - 04 = Técnico /Técnica Superior de Formación Profesional o título equivalente
    - 05 = Técnico/Técnica Superior de Artes plásticas y diseño o título equivalente
    - 06 = Técnico Deportivo/Técnica Deportiva Superior o título equivalente
  - Si se trata de un o una estudiante que accede por primera vez al SUE con forma de acceso distinta de 01 y 03:
    - 88 = No aplica

    - ◇ 99 = No consta
  - Si se trata de un o una estudiante que no accede por primera vez al SUE:
    - ◇ 88 = No aplica
    - ◇ 99 = No consta

- **Especialidad del estudio de acceso a este Grado universitario [EspecialidadAcceso].** Si en la variable "Estudio de acceso a este Grado universitario" se ha consignado un número distinto a 01-04, el valor es
  - 8888 = No aplica
- **Municipio del centro en el que cursó el último año del estudio que le da acceso a este Grado universitario [MunicipioCentroSec].** Código del municipio en el que está ubicado el centro donde cursó el último año del estudio que le da acceso.
  - 80000 = Municipio perteneciente al EEES
  - 80001 = Municipio no perteneciente al EEES
  - 88888 = No aplica
  - 99999 = No consta
- **Último año que cursó el estudio que le da acceso a este Grado universitario [AnioFinEstudioAcceso].** Se consignará el último año natural del curso académico
  - 9999 = No consta
- **País en que cursó el último año del estudio que le da acceso a este Grado universitario [PaisFinEstudioAcceso].** Código del país en que cursó el último año de estudio.
- **Naturaleza del centro en el que cursó el estudio que le da acceso a este Grado universitario [NaturalezaCentroSec].**
  - Si se trata de una o un estudiante que accede por primera vez al SUE:
    - ◇ 1 = Centro Público
    - ◇ 2 = Centro Privado
    - ◇ 3 = Centro Privado Concertado
    - ◇ 9 = No consta
  - En otro caso:
    - ◇ 8 = No aplica
- **Accede por primera vez al SUE este curso [NuevoSUE].**
  - 0 = No

- 1 = Sí

- **Año de acceso al SUE [AnioAccesoSUE].** Los cuatro dígitos del primer año de los dos años naturales que compone el curso académico en el que el estudiante ingresó por primera vez en el Sistema Universitario Español.
  - 9999 = No consta
- **Nota de admisión al estudio que cursa actualmente [NotaAdmision]**. La nota global de admisión con la que el estudiante ha accedido al título de grado universitario que cursa actualmente. En caso de que la forma de acceso sea por poseer otro título superior se utilizará el baremo propio de dicho título. El formato del campo será nnnnn donde las tres últimas posiciones será la parte decimal de la nota. La parte entera debe estar separada de la decimal por un punto.
  - 88.888 = No aplica

- ***Tabla de Avance***. Aparecen las siguientes variables:

  - **Universidad**. Código de la universidad en la que se imparte el Grado.
  - **Centro**. Código del centro en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Municipio**. Código del municipio en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Campus**. Código del campus en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Titulación**. Código del Grado en el que está matriculado el estudiante.
  - **Tipo de documento identificativo [TipoDocumento]**. Tipo de documento que presenta el estudiante. Valores:
    - 1 = NIF
    - 2 = NIE
    - 3 = Pasaporte
    - 4 = Otros
  - **Nº de documento identificativo [NumeroDocumento]**. Número del documento que presenta el estudiante. El formato será:
    - NIF = nnnnnnnnL(donde nnnnnnn serán los 8 número y L la letra final que corresponda)
    - NIE = AnnnnnnnL (A + 7 números + letra final que corresponda)

  - Pasaporte u otros. El formato del pasaporte no se validará.
- **Año de nacimiento** [**AnioNacimiento**]. Los cuatro dígitos del año de nacimiento
- **Nacionalidad** [**PaisNacionalidad**]**.** Código del país de nacionalidad.
- **Sexo.**
  - H = Hombre
  - M = Mujer
- **Modalidad reconocida de dedicación al estudio** [**DedicacionEstudio**]**.**
  - 1 = Modalidad tiempo completo
  - 2 = Modalidad tiempo parcial
  - 3 = Otros

- ***Tabla de Rendimiento***. Incluye las siguientes variables.
  - **Universidad**. Código de la universidad en la que se imparte el Grado.
  - **Centro**. Código del centro en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Municipio**. Código del municipio en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Campus**. Código del campus en el que se imparte el Grado en el que está matriculado el estudiante.
  - **Titulación**. Código del Grado en el que está matriculado el estudiante.
  - **Tipo de documento identificativo** [**TipoDocIndentidad**]. Tipo de documento que presenta el estudiante. Valores:
    - 1 = NIF
    - 2 = NIE
    - 3 = Pasaporte
    - 4 = Otros
  - **Nº de documento identificativo** [**NumeroDocumento**]. Número del documento que presenta el estudiante. El formato será:
    - NIF = nnnnnnnnL(donde nnnnnnn serán los 8 número y L la letra final que corresponda)

- NIE = AnnnnnnnL (A + 7 números + letra final que corresponda)
- Pasaporte u otros. El formato del pasaporte no se validará.

- **Año de nacimiento [AnioNacimiento]**. Los cuatro dígitos del año de nacimiento
- **Nacionalidad [PaisNacionalidad].** Código del país de nacionalidad.
- **Sexo.**
  - H = Hombre
  - M = Mujer
- **Modalidad reconocida de dedicación al estudio [DedicacionEstudio].**
  - 1 = Modalidad tiempo completo
  - 2 = Modalidad tiempo parcial
  - 3 = Otros

- **País del domicilio familiar de residencia habitual [PaisResidencia].** Se consignará el país donde el o la estudiante reside habitualmente con su familia.
- **Municipio del domicilio familiar de residencia habitual [MunicipioResidencia].** Se consignará el municipio donde el o la estudiante reside habitualmente con su familia.
  - 80000 = Municipio perteneciente al EEES
  - 80001 = Municipio no perteneciente al EEES
  - 88888 = No aplica
  - 99999 = No consta
- **Municipio de residencia del estudiante durante el curso [MunicipioResidenciaCurso].** Código del municipio de residencia del o de la estudiante durante el curso académico.
  - 80000 = Municipio perteneciente al EEES
  - 80001 = Municipio no perteneciente al EEES
  - 88888 = No aplica
  - 99999 = No consta
- **Trabajó el estudiante durante el curso anterior [TrabajoAnterior].** Si realizó algún trabajo o actividad remunerada, ya sea en la universidad o fuera de ella.

- 0 = No realizó ningún trabajo o actividad remunerada
- 1 = Trabajo esporádico (durante menos de tres meses)
- 2 = Trabajo a jornada Parcial (durante más de tres meses)
- 3 = Trabajo a jornada Completa. (durante más de tres meses)

- **Año de acceso al SUE [AnioAccesoSUE].** Los cuatro dígitos del primer año de los dos años naturales que compone el curso académico en el que el estudiante ingresó por primera vez en el Sistema Universitario Español.
  - 9999 = No consta
- **Año de acceso a este Grado universitario [AnioAccesoTitulacion].** Los cuatro dígitos del primer año de los dos años naturales que compone el curso académico en el que el estudiante ingresó por primera vez en este Grado universitario
- **Créditos ordinarios matriculados en el curso académico de referencia [CreditosMat].** Número total de créditos ordinarios en los que el estudiante se ha matriculado a lo largo del curso académico de referencia. No se incluirán los créditos que se hayan matriculado con objeto de reconocimiento. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios matriculados en el curso académico de referencia por primera vez [CreditosMat1].** Número de créditos ordinarios en los que el estudiante se ha matriculado por primera vez en el curso académico de referencia. No se incluirán los créditos que se hayan matriculado con objeto de reconocimiento. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios matriculados en el curso académico de referencia por segunda vez [CreditosMat2].** Número de créditos ordinarios en los que el estudiante se ha matriculado por segunda vez en el curso académico de referencia. No se incluirán los créditos que se hayan matriculado con objeto de reconocimiento. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios matriculados en el curso académico de referencia por tercera o más veces [CreditosMat3].** Número de créditos ordinarios en los que el estudiante se ha matriculado por tercera vez o más en el curso académico de referencia. No se incluirán los créditos que se hayan

matriculado con objeto de reconocimiento. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.

- **Créditos ordinarios superados en el curso académico de referencia [CreditosSupCurso].** Número de créditos ordinarios que el estudiante ha superado en el curso académico de referencia, habiéndolos matriculado. La suma de créditos ordinarios superados y créditos ordinarios no superados debe ser igual al número de créditos ordinarios matriculados en el curso académico de referencia. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios no superados en el curso académico de referencia [CreditosNoSupCurso].** Número de créditos ordinarios que el estudiante no ha superado en el curso académico de referencia, habiéndolos matriculado. La suma de créditos ordinarios superados y créditos ordinarios no superados debe ser igual al número de créditos ordinarios matriculados en el curso académico de referencia. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos reconocidos en el curso académico de referencia [CreditosRecCurso].** Número de créditos que el estudiante ha solicitado y le han sido reconocidos en el curso académico de referencia. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios presentados en el curso académico de referencia [CreditosPreCurso].** Número de créditos ordinarios que habiendo sido matriculados en el curso académico de referencia el estudiante se ha presentado a evaluación durante este curso. La suma de créditos ordinarios presentados y créditos ordinarios no presentados debe ser igual al número de créditos ordinarios matriculados en el curso académico de referencia. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios no presentados en el curso académico de referencia[CreditosNoPreCurso].** Número de créditos ordinarios que habiendo sido matriculados en el curso académico de referencia el estudiante no se ha presentado a evaluación durante este curso. La suma de créditos ordinarios presentados y créditos ordinarios no presentados debe ser igual al número de créditos ordinarios matriculados en el curso académico de referencia. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.

- **Créditos ordinarios matriculados desde el inicio de este Grado universitario [CreditosMatInicio].** Número total de créditos que el estudiante se ha matriculado desde que inició el Grado universitario. No incluye los créditos reconocidos ni los créditos transferidos en cada curso académico. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios superados desde el inicio de este Grado universitario [CreditosSupInicio].** Número total de créditos que el estudiante ha superado desde que inició el estudio que cursa actualmente. No incluye los créditos reconocidos ni los créditos transferidos en cada curso académico. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos ordinarios presentados desde el inicio de este Grado universitario [CreditosPreInicio].** Número total de créditos ordinarios que, habiendo sido matriculados, el estudiante se ha presentado a evaluación desde que inició este Grado universitario. No incluye los créditos reconocidos ni los créditos transferidos en cada curso académico. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Créditos reconocidos desde el inicio de este Grado universitario [CreditosRecInicio].** Número total de créditos que le han sido reconocido al estudiante desde que inició este Grado universitario. Las dos últimas posiciones corresponden a la parte decimal del número de créditos. La parte entera debe estar separada de la decimal por un punto.
- **Total de créditos transferidos [TotalCreditosTransferidos].** Número de créditos obtenidos en enseñanzas oficiales cursadas con anterioridad que forman parte del expediente académico del estudiante y que no han conducido a la obtención de un título. La parte entera debe estar separada de la decimal por un punto.
- **Titulado en el curso de referencia [TituladoCursoRef].** El estudiante será titulado/a si ha completado los créditos suficientes para obtener el título, con independencia de haber abonado las tasas correspondientes.
  - 0 = No
  - 1 = Sí
- **Nota media del expediente académico (si el estudiante es titulado/a en el curso de referencia) [NotaMediaExpediente].** Nota media del expediente académico del estudio en el que se titula el o la estudiante en el curso de referencia. La forma de cálculo, según el RD 1125/2003, será la

suma de los créditos obtenidos por el alumno o por la alumna multiplicados cada uno de ellos por el valor de las calificaciones que correspondan, y dividida por el número de créditos totales obtenidos por el alumno o por la alumna. Los resultados obtenidos por el alumno o la alumna en cada una de las materias del plan de estudios se calificarán en función de la siguiente escala numérica de 0 a 10, con expresión de un decimal. Los créditos obtenidos por reconocimiento de créditos correspondientes a actividades formativas no integradas en el plan de estudios NO se computarán a efectos de cómputo de la media del expediente académico. La parte entera debe estar separada de la decimal por un punto. Si el estudiante no ha sido titulado este curso consignar:

- 88.88 = No aplica.

3.5.2. Datos objetivo

El conjunto inicial de datos no es el más indicado para trabajar, dado que hay muchos alumnos en las distintas tablas que no son de nuestro interés, por lo que ha sido necesario realizar una serie de filtrados para trabajar con el subconjunto de datos deseado. Los filtros han sido aplicados por años, de modo que se ha obtenido el subconjunto deseado de cada año. Hemos definido tres tipos de filtrado distintos:

- **Filtrado Tipo I**: Este filtrado consiste en quedarse con aquellos alumnos que aparecen en la tabla de acceso y rendimiento de un año N y que su año de acceso al SUE (Sistema Universitario Español) sea también N.
- **Filtrado Tipo II**: Este filtrado consiste en quedarse con aquellos alumnos que aparecen en la tabla de acceso de un año N.
- **Filtrado Tipo III**: Este filtrado consiste en quedarse con aquellos alumnos que aparecen en la tabla de acceso y rendimiento de un año N.

Se ha aplicado el filtro de tipo I para los años 2009-2010 y 2010-2011, el filtrado de tipo II para los años 2011-2012 y 2012-2013 y el filtrado de tipo III también para los años 2011-2012 y 2012-2013.

3.5.2.1. Informe de Calidad

Se ha constatado un problema en la calidad de los datos recopilados: muchos campos no se han rellenado a un valor concreto, dejándose a NULL. Se considera relevante tratar de concienciar al personal de admisión en que la recogida de información sobre matrícula en todos los campos es esencial para poder elaborar informes de calidad.

Para realizar el informe de calidad se han dividido los datos de dos maneras:

- **Conjunto de datos I**: Consiste en unir en una tabla las instancias del curso 2009-2010 con el filtrado de tipo I, del curso 2010-2011 con el filtrado de tipo I, del curso 2011-2012 con el filtrado de tipo II y del curso 2012-2013 con el filtrado de tipo II. Este conjunto de datos tiene en total 983 instancias.
- **Conjunto de datos II**: Consiste en unir en una tabla las instancias del curso 2009-2010 con el filtrado de tipo I, del curso 2010-2011 con el filtrado de tipo I, del curso 2011-2012 con el filtrado de tipo III y del curso 2012-2013 con el filtrado de tipo III. Este conjunto de datos tiene en total 960 instancias.

Se ha realizado un estudio sobre el número de valores NULL de las siguientes variables:

NivelEstudioPadre, NivelEstudioMadre, OcupaciónPadre, Ocupación Madre.

En el conjunto de datos I:

- Hay 165 de 983 que tienen todos esos a NULL, es decir, un 16,78 %, que se puede ver en la figura 3.1
- Hay 436 de 983 que tiene algún atributo de esos a NULL, es decir, un 44,35 %. Están incluidos los anteriores. Se puede ver en la figura 3.2

En el conjunto de datos II:

- Hay 151 de 960 que tienen todos esos a NULL, es decir, un 15,73 %, que se puede ver en la figura 3.1
- Hay 423 de 960 que tiene algún atributo de esos a NULL, es decir, un 44,06 %. Están incluidos los anteriores. Se puede ver en la figura 3.2

En la figura 3.1 se puede ver el porcentaje de alumnos que tienen todas las variasbles a NULL tanto del conjunto de datos I como el conjunto de datos II, dado que los porcentajes son prácticamente iguales. Igualmente, 3.2 se puede ver el porcentaje de alumnos que tienen alguno de estos datos a NULL de ambos conjuntos.

NaturalezaCentroSec

En el conjunto de datos I:

- Hay 164 de 983 que tienen todos esos a NULL, es decir, un 16,68 %.

En el conjunto de datos II:

Figura 3.1: Alumnos del conjunto de datos I con NivelEstudioPadre, NivelEstudioMadre, OcupaciónPadre y OcupaciónMadre a NULL

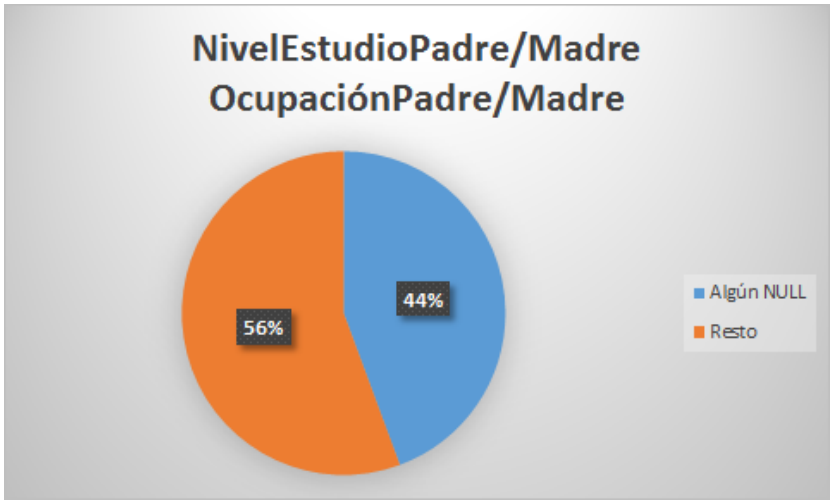

Figura 3.2: Alumnos del conjunto de datos I con NivelEstudioPadre, NivelEstudioMadre, OcupaciónPadre o OcupaciónMadre (al menos uno de ellos) a NULL

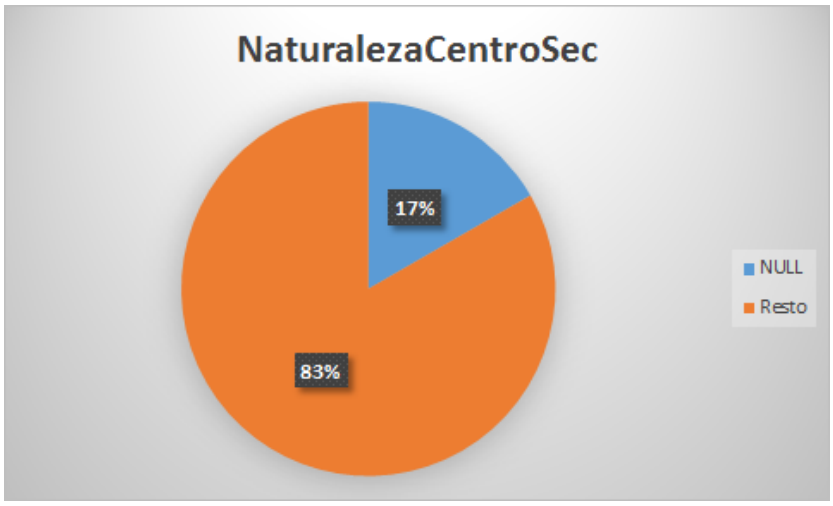

Figura 3.3: Alumnos del conjunto de datos I con NaturalezaCentroSec a NULL

- Hay 164 de 960 que tienen todos esos a NULL, es decir, un 17,08 %.

En la figura 3.3 se puede observar el gráfico con el porcentaje de alumnos que tienen la variable NaturalezaCentroSec a NULL, tanto del conjunto de datos I como el conjunto de datos II, dado que los porcentajes son muy similares.

MunicipioCentroSec

En el conjunto de datos I:

- Hay 345 de 983 que tienen este campo a NULL, es decir, un 35,10 %.

En el conjunto de datos II:

- Hay 340 de 960 que tienen este campo a NULL, es decir, un 35,42 %.

En la figura 3.4 se puede ver el porcentaje de alumnos que tienen la variable MunicipioCentroSec a NULL, tanto del conjunto de datos I como el conjunto de datos II, dado que los porcentajes son muy similares.

AnioNacimiento y AnioAccesoSUE

En el conjunto de datos I:

- Hay 36 de 983 que han entrado con menos de 15 años (o antes de nacer), es decir, un 3,66 %.

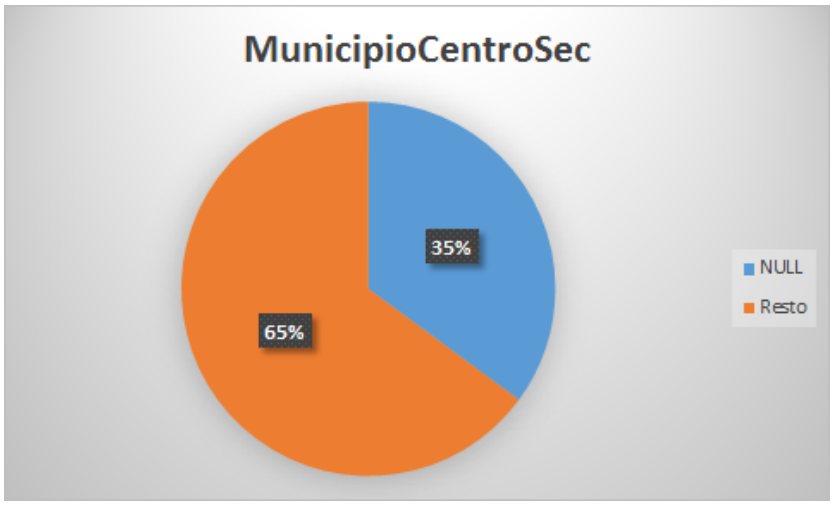

Figura 3.4: Alumnos del conjunto de datos I con MunicipioCentroSec a NULL

- Hay 27 de 983 que han entrado antes de nacer, es decir, un 2,74 %.

En el conjunto de datos II:

- Hay 31 de 960 que han entrado con menos de 15 años (o antes de nacer), es decir, un 3,23 %.
- Hay 22 de 960 que han entrado antes de nacer, es decir, un 2,29 %.

NotaAdmision

En el conjunto de datos I:

- Hay 46 de 983 de los que no se dispone de nota de entrada en selectividad, es decir, un 4,68 %.

En el conjunto de datos II:

- Hay 46 de 960 de los que no se dispone de nota de entrada en selectividad, es decir, un 4,79 %.

3.5.3. Elección del conjunto de datos

A partir de este punto y a lo largo del resto del documento se utilizará el **conjunto de datos I**, que tiene 983 instancias de alumnos. Esto no es debido al estudio de la calidad de las variables, sino al contexto.

En el conjunto de datos I están incluidos todos los alumnos que aparecen en las tablas de acceso, los cuales, según la descripción de los ficheros del SIIU, son aquellos que se han matriculado por primera vez en el grado del curso académico del año correspondiente.

En cambio, en el conjunto de datos II, con el fin de excluir aquellos alumnos que no han acabado ni siquiera el primer año, se ha podido quitar instancias importantes y con ello sobreajustar los modelos que saldrían de esos datos.

3.5.4. Preparación de los datos

Dado que no existe un atributo abandono, generamos la variable a predecir, teniendo en cuenta la semántica del problema. Para calcular el abandono observamos cuántos alumnos aparecen en la tabla del año N (independientemente del tipo de filtrado aplicado) y en la tabla de Rendimiento del año N+1. Si por algún motivo no existiese la tabla de Rendimiento del año N+1 (por ejemplo, del año en curso), se utilizaría la tabla de Avance del año N+1.

Con estos datos, han abandonado 222 alumnos de 983, es decir, un **22.5 %** de abandono entre los cursos 2009-2010 y 2012-2013.

También ha sido necesario extraer variables implícitas con el fin de mejorar nuestro modelo. Las variables obtenidas son las siguientes:

- **Rezagado:**
  - 0 → ha entrado a la universidad con 18 años.
  - 1 → ha entrado a la universidad con 19 años.
  - 2 → ha entrado a la universidad con 20 años o más.
- **Español:**
  - 0 → es español.
  - 1 → no es español.
- **CentroSecMadrid:**
  - 0 → el centro de secundaria pertenece a Madrid.
  - 1 → el centro de secundaria no pertenece a Madrid.
- **TrabajaEstudiante:**
  - 0 → no trabaja el estudiante.
  - 1 → el estudiante tiene trabajo.
- **Edad**

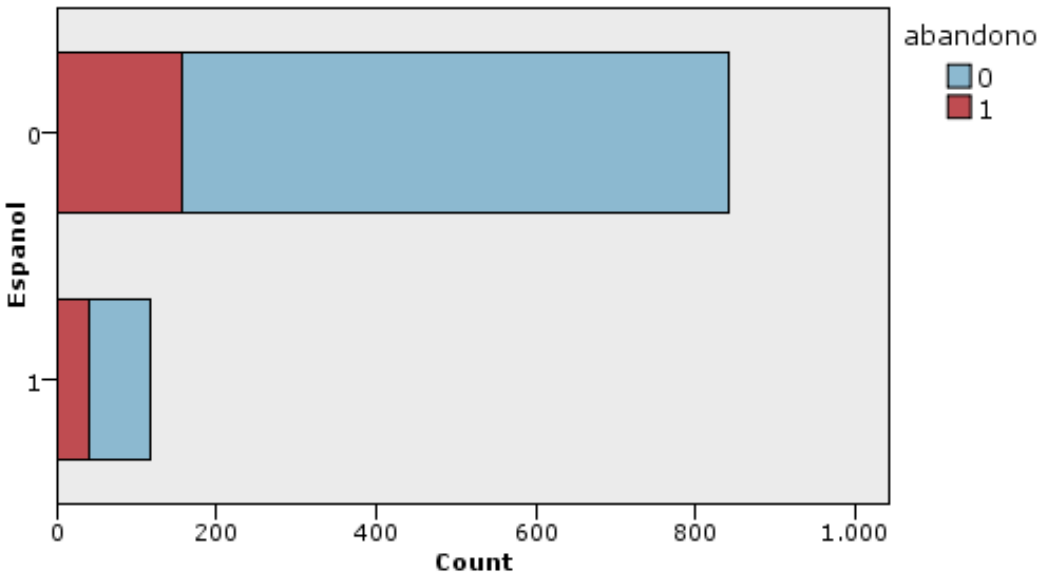

Figura 3.5: Español sin normalizar

- **Rango Nota:** La variable NotaAdmisión es una variable continua que se discretizó de la siguiente manera:
  - 0 → entre 5 y 6.
  - 1 → entre 6 y 7.
  - 2 → entre 7 y 8.
  - 2 → entre 8 y 14.

3.5.5. Análisis de las variables derivadas

Una vez obtenidas las variables deseadas, podemos estudiar qué relación hay entre dichas variables y el abandono.

En las siguientes figuras veremos en rojo (1) a los alumnos que abandonan y en azul (0) a aquellos que no abandonan.

3.5.5.1. Abandono con Español

En las figuras 3.5 y 3.6 se puede determinar que aquellos que no son de nacionalidad española tienen mayor porcentaje de abandono.

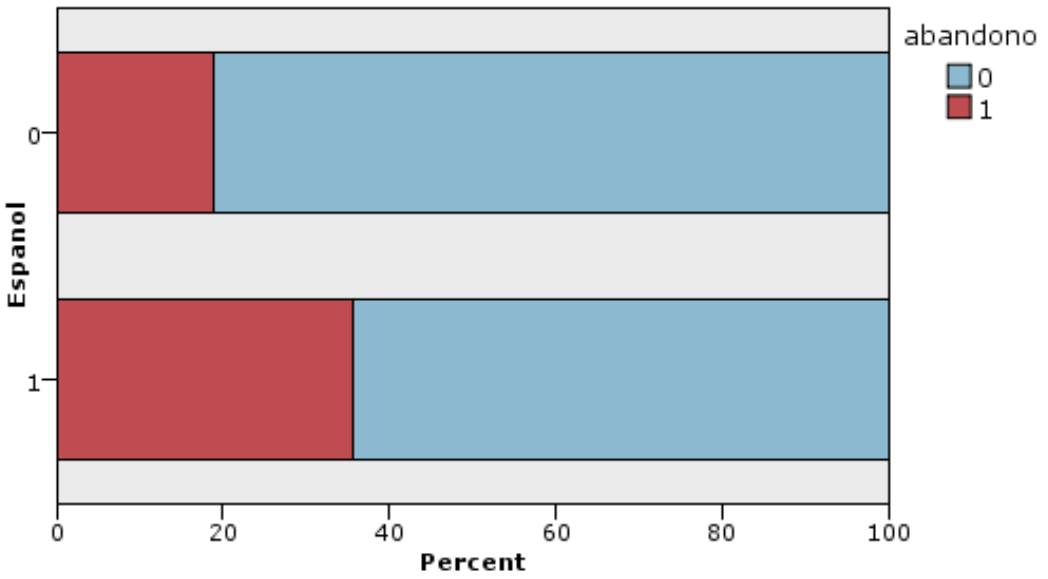

Figura 3.6: Español normalizado

3.5.5.2. Abandono con Rezagado

En las figuras 3.7 y 3.8 se puede ver de manera clara cómo aquello que han entrado a SUE con 20 años o más tienen mayor porcentaje de abandono que los que entraron con 18 o 19. Además, aquellos que entraron con 18 tienen sensiblemente menor porcentaje de abandono que los que entraron con 19.

3.5.5.3. Abandono con RangoNota

En las figuras 3.9 y 3.10 podemos observar que la diferencia de nota de acceso es un atributo bastante representativo, puesto que en el rango 3 (nota entre 8 y 14) el descenso del porcentaje de abandono es muy grande. Cabe destacar que en estas figuras se han eliminado aquellos alumnos que por diversos motivos su nota de acceso es NULL.

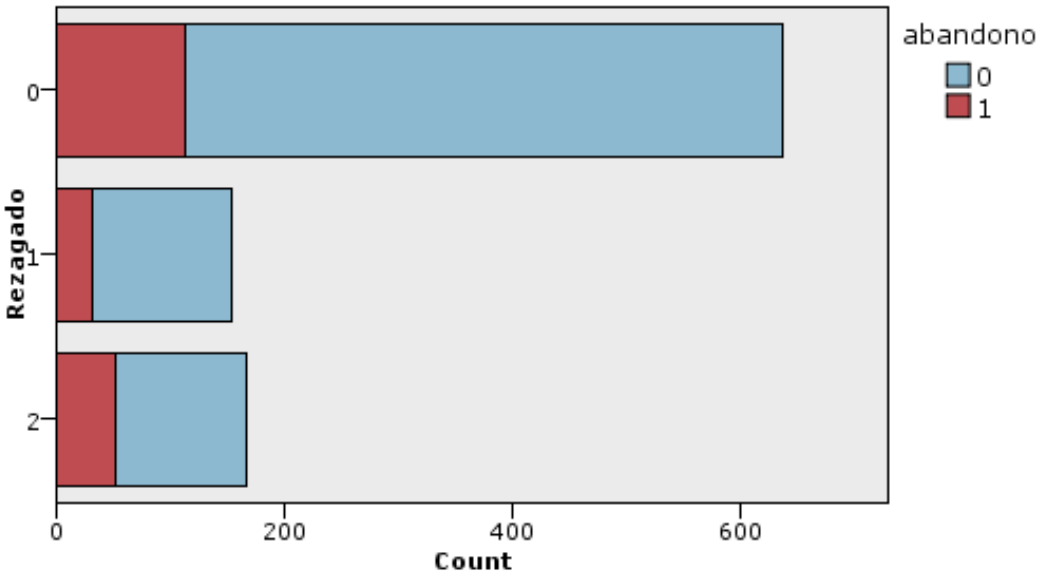

Figura 3.7: Rezagado sin normalizar

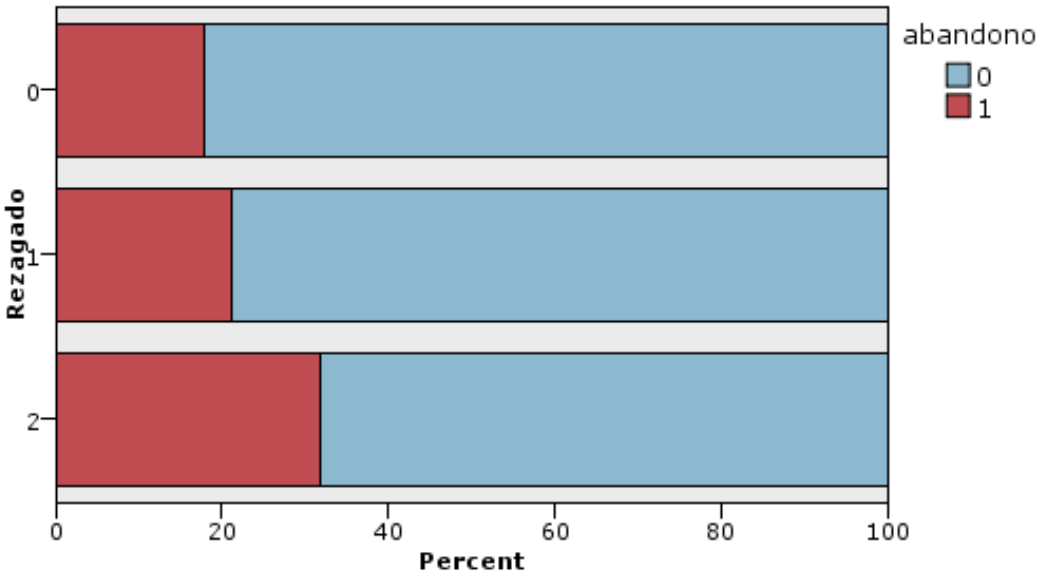

Figura 3.8: Rezagado normalizado

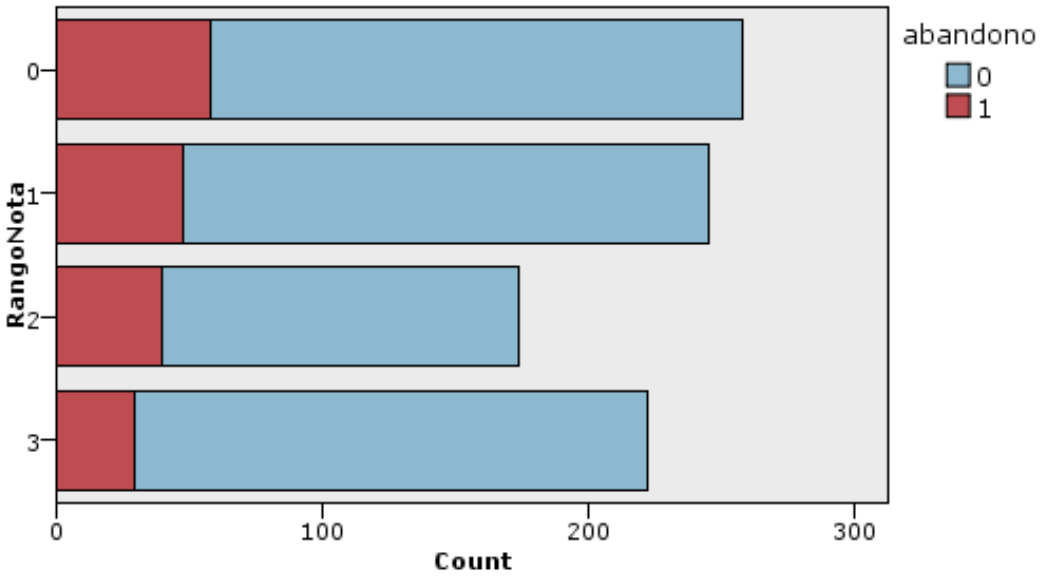

Figura 3.9: RangoNota sin normalizar

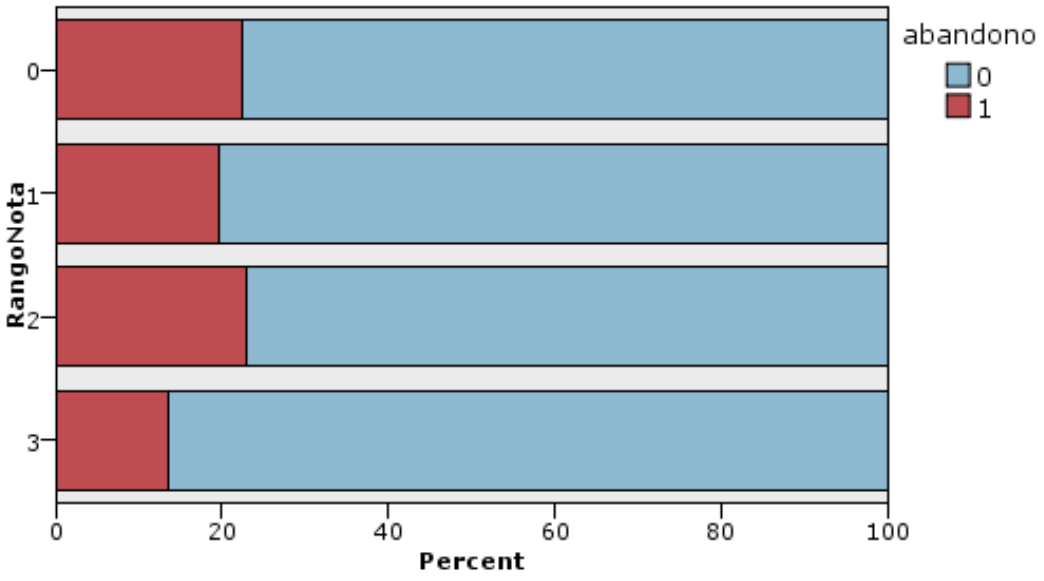

Figura 3.10: RangoNota normalizado

Capítulo 4

Modelización y evaluación

4.1. Fórmula para la comparación de modelos

Para poder elegir el mejor modelos de *Data Mining*, ha sido necesario establecer una fórmula que haga posible la comparación entre ellos, dado que el fin no es solo tener un gran porcentaje de aciertos, sino que además de eso, hay que minimizar aquellos que el modelo predice como no abandono pero que finalmente abandonan. Para ello, se ha decidido que la fórmula actúe sobre la matriz de confusión resultante.

True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

Tabla 4.1: Matriz de confusión

De este modo, dada la matriz de confusión descrita en el cuadro 4.1, se ha definido la siguiente fórmula para comparar modelos:

$$0{,}7 * \frac{TN}{FP + TN} + 0{,}3 * \frac{TP}{FN + TP}$$

La fórmula está acotada entre 0 y 1, siendo 0 el peor resultado y 1 el mejor.

Un modelo será mejor que otro si el resultado de aplicar la fórmula sobre la matriz de confusión obtenida es mayor que el otro.

4.2. Variables

Las variables disponibles son las siguientes:

- Universidad

- Centro
- Municipio
- Campus
- Titulación
- Tipo Documento Identidad
- Año de Nacimiento
- País Nacionalidad
- Sexo
- Dedicación Estudio
- Ocupación Estudiante
- Familia Numerosa
- Nivel Estudio Padre
- Nivel Estudio Madre
- Ocupación Padre
- Ocupación Madre
- Forma Admisión
- Estudio Acceso
- Especialidad Acceso
- País Fin de Estudio Acceso
- Naturaleza Centro Secundaria
- Nuevo SUE
- Año Acceso SUE
- Nota de Admisión
- Rezagado
- Centro Secundaria Madrid

- Trabaja Estudiante
- Español
- Edad
- Rango Nota

4.3. Pruebas con Modelos

4.3.1. Primer contacto

Una vez decidida la manera de comparar los distintos modelos, es momento de comparar los mismos. Para ello, se ha realizado la técnica de validación cruzada, con la técnica de leave-one-out, dado que es el mejor método de validación cruzada [Kohavi95] siempre y cuando se tengan más de 60 instancias.

En primer lugar se han utilizado todas las variables disponibles, convertidas en String (excepto NotaAdmisión, AnioNacimiento y AnioAccesoSUE), de modo que los algoritmos de tipo árbol no realicen cálculos a través de las distancias entre las distintas variables (dado que en este caso los números no representan distancias). Las pruebas se han realizado con los distintos algoritmos que se pueden encontrar en la tabla 4.2.

Algoritmo	Resultado Fórmula
SMO	0,6985
Naive Bayes Tree	0,7125
Random Forest	0,7128
Logistic	0,7078
KNIME Tree	0,6991
C4.5	0,7074

Tabla 4.2: Comparación de algoritmos utilizando todas las variables disponible

Los malos resultados obtenidos pueden deberse a varios motivos. Algunos de ellos son los siguientes:

- La elección de variables puede no ser la adecuada. Hay muchas variables que se no tienen importancia a nivel semántico. Puede que con una mejor elección de variables, el resultado mejore.
- De los 983 alumnos que tenemos en la tabla, hay 761 que no abandonan y 222 que sí que lo hacen, lo que significa que la clase no está ajustada. Una manera de mejorar el modelo sería igualar el número de alumnos que no abandonan con los

que sí abandonan, bien generando artificialmente alumnos que abandonan o bien filtrando alumnos que no abandonan.

- Por último, cabe la posibilidad de que con estos datos no exista un buen modelo capaz de predecir el abandono.

4.3.2. Reelección de variables

Lo primero que vamos a hacer para intentar mejorar el resultado será escoger las variables que más sentido tienen en este problema a nivel semántico. Además, se eliminarán también aquellas variables que estén a NULL en más de la mitad de los alumnos. Las variables elegidas siguiendo lo comentado anteriormente son las siguientes:

- País Nacionalidad
- Sexo
- Dedicación Estudio
- Familia Numerosa
- Nivel Estudio Padre
- Nivel Estudio Madre
- Ocupación Padre
- Ocupación Madre
- Forma Admisión
- País Fin de Estudio Acceso
- Naturaleza Centro Secundaria
- Año Acceso SUE
- Nota de Admisión
- Rezagado
- Centro Secundaria Madrid
- Edad

Con estas variables convertidas a String (excepto AnioAccesoSUE, NotaAdmisión y Edad) por los problemas que tienen los algoritmos de tipo árbol, se han obtenido los resultados que aparecen en la tabla 4.3. Como se puede apreciar en dicha tabla, los resultados han sido bastante malos.

Algoritmo	Resultado Fórmula
SMO	0,7068
Naive Bayes Tree	0,7069
Random Forest	0,7033
Logistic	0,7039
KNIME Tree	0,6814
C4.5	0,7102

Tabla 4.3: Comparación de algoritmos utilizando una selección de variables a nivel semántico

4.3.3. Ajuste de la distribución de la clase

Existen dos maneras para ajustar la distribución de la clase [Rahman13]:

- **Oversampling** (sobremuestreo): Consiste en generar de manera artificial instancias de una de las clases con el fin de ajustar la distribución del conjunto de datos. Esta generación de instancias se realiza con distintos algoritmos como K-Means.
- **Subsampling** o **sampling** (muestreo): Consiste en eliminar o descargar instancias de la clase predominante, con el fin de ajustar la distribución del conjunto de datos. Esta eliminación se suele realizar de manera aleatoria y estratificada.

Dado que el tiempo disponible para realizar el proyecto es limitado, solo se realizará una de las técnicas para ajustar la distribución de la clase: Oversampling. El conjunto de datos, después del ajuste realizado, contiene 1522 instancias. Con este conjunto de datos y con todas las variables disponibles, se han obtenido los resultados que se pueden ver en la tablas 4.4.

Algoritmo	Resultado Fórmula
SMO	0,6318
Naive Bayes Tree	0,7434
Random Forest	**0,8308**
Logistic	0,6457
KNIME Tree	0,7477
C4.5	0,7503

Tabla 4.4: Comparación de algoritmos utilizando la técnica de oversampling y todas las variables disponibles

Además, se han realizado también los modelos con el conjunto de variables con significado semántico, que se pueden encontrar en la tabla 4.5. En ambas tablas se

aprecia también como el algoritmo de Random Forest es el que destaca del resto. La matriz de confusión resultante de ambos método se puede encontrar en la tabla 4.6. También en dicha tabla se puede observar que ambos conjuntos de matrices cumplen con los criterios de éxito, es decir, el *accuracy* (porcentaje de bien clasificados) está en torno al 80 %±2 % y los False Negative (aquellos clasificados como *no abandono* que finalmente abandonan) menor que el 5 %.

Algoritmo	Resultado Fórmula
SMO	0,6525
Naive Bayes Tree	0,7409
Random Forest	**0,7595**
Logistic	0,6315
KNIME Tree	0,7464
C4.5	0,6625

Tabla 4.5: Comparación de algoritmos utilizando la técnica de oversampling y una selección de variables a nivel semántico

Conjunto de variables	Matriz de confusión		Fórmula	Accuracy	False Negative
Todas las variables	599	162	0,8308	86,01 %	3,35 %
	51	710			
Selección de variables	524	237	0,7595	80,68 %	3,74 %
	57	704			

Tabla 4.6: Comparación de los modelos con oversampling y Random Forest a través de la selección de variables

Por último, en la tabla 4.7 se pueden encontrar las variables más importantes para el modelo tanto con todas las variables como con la selección semántica.

4.4. Elección del Modelo

Como hemos visto en las secciones anteriores, el mejor modelo es aquel que contiene el algoritmo Random Forest, después de efectuar la técnica de oversampling para ajustar la distribución de la clase. Aún así, no se debe afirmar con rotundidad que este modelo es el mejor, puesto que es posible que no se ajuste bien a los datos en un futuro debido a los métodos para reducir el abandono que se aplicarán en años posteriores.

Conjunto de variables	Variables más Importantes
Todas las variables	PaísNacionalidad FormaAdmisión TipoDocIdentidad Español PaísFinEstudioAcceso NotaAdmisión Edad
Selección de variables	País Nacionalidad Forma Admisión Edad PaísFinEstudioAcceso NotaAdmisión Rezagado AñoAccesoSUE

Tabla 4.7: Mejores variables de los distintos Conjuntos de Datos

A pesar de ello, dicho método es el elegido para dar solución al problema puesto y queda a elección del usuario las variables a utilizar, dado que ambos métodos cumplen los criterios de éxito.

Capítulo 5

Sistematización

Uno de los fines de este proyecto es que se pueda repetir en un futuro. Para ello, se ha intentado documentar cada paso llevado a cabo, con el fin de sistematizarlo.

5.1. Obtención y manejo de los ficheros XML

El primer paso y el más importante es obtener los ficheros del SIIU. Para la realización de este proyecto, se han obtenido los ficheros de rendimiento y acceso de los cursos 2009-2010, 2010-2011, 2011-2012 y 2012-2013 y, además, el fichero de avance del curso 2013-2014.

Para cargar los ficheros a la base de datos SQL, hay que sustituir "<Registro>" y "$< Registro >$" por "<row>" y "</row>", respectivamente. Además, para eliminar los saltos de línea en los campos, hay que sustituir "$< ([A-Z][a-zA-Z]*) > [\r\n]+$" por "$< \1 >$" (en algunos editores de texto hay que activar la opción "expresión regular"). En este proyecto se ha utilizado phpMyAdmin y para cargar ficheros xml hay que eliminar además la cabecera y cierre "<Fichero>" y "</Fichero>".

Una vez hecho esto, ya se pueden cargar los ficheros a la base de datos, que la llamaremos "abandonoupm". Las tablas tendrán los nombres de la siguiente manera: "nombreCurso", de modo que los nombres serán "acc" para las tablas de Acceso, "rend" para las tablas de Rendimiento y "ava" para las tablas de Avance. El curso vendrá dado con las dos últimas cifras. Por ejemplo, el fichero de Acceso del curso 2009-2010 se guardará en la base de datos como "acc09-10".

5.1.1. Cargar un fichero de acceso

Para cargar un fichero de acceso, basta con introducir la siguiente sentencia SQL. Se debe sustituir "NOMBRE" por el nombre que se le quiera dar a la tabla y "PATH" por el path completo donde se encuentra el fichero XML.

```
CREATE TABLE `NOMBRE`
(Universidad char(50),
        Centro char(50),
        Municipio  char(50),
        Campus char(50),
        Titulacion char(50),
        TipoDocIdentidad char(50),
        NumeroDocumento char(50),
        AnioNacimiento char(50),
        PaisNacionalidad char(50),
        Sexo char(50),
        DedicacionEstudio char(50),
        OcupacionEstudiante char(50),
        FamiliaNum char(50),
        NivelEstudioPadre char(50),
        NivelEstudioMadre char(50),
        OcupacionPadre char(50),
        OcupacionMadre char(50),
        FormaAdmision char(50),
        EstudioAcceso char(50),
        EspecialidadAcceso char(50),
        MunicipioCentroSec char(50),
        PaisFinEstudioAcceso char(50),
        NaturalezaCentroSec char(50),
        NuevoSUE char(50),
        AnioAccesoSUE char(50),
        NotaAdmision char(50));

load xml infile "PATH" into table `NOMBRE`
```

5.1.2. Cargar un fichero de rendimiento

Para cargar un fichero de rendimiento, basta con introducir la siguiente sentencia SQL. Se debe sustituir "NOMBRE" por el nombre que se le quiera dar a la tabla y "PATH" por el path completo donde se encuentra el fichero XML.

```
CREATE TABLE `NOMBRE`
(Universidad char(50),
        Centro char(50),
        Municipio  char(50),
        Campus char(50),
```

```
        Titulacion char(50),
        TipoDocIdentidad char(50),
        NumeroDocumento char(50),
        AnioNacimiento char(50),
        PaisNacionalidad char(50),
        Sexo char(50),
        DedicacionEstudio char(50),
        PaisResidencia char(50),
        MunicipioResidencia char(50),
        MunicipioResidenciaCurso char(50),
        TrabajoAnterior char(50),
        AnioAccesoSUE char(50),
        AnioAccesoTitulacion char(50),
        CreditosMat char(50),
        CreditosMat1 char(50),
        CreditosMat2 char(50),
        CreditosMat3 char(50),
        CreditosSupCurso char(50),
        CreditosNoSupCurso char(50),
        CreditosRecCurso char(50),
        CreditosPreCurso char(50),
        CreditosNoPreCurso char(50),
        CreditosMatInicio char(50),
        CreditosSupInicio char(50),
        CreditosPreInicio char(50),
        CreditosRecInicio char(50),
        TotalCreditosTransferidos char(50),
        TituladoCursoRef char(50),
        NotaMediaExpediente char(50));

load xml infile "PATH" into table `NOMBRE`
```

5.1.3. Cargar un fichero de Avance

Para cargar un fichero de rendimiento, basta con introducir la misma sentencia SQL que en 5.1.2. Se debe sustituir "NOMBRE" por el nombre que se le quiera dar a la tabla y "PATH" por el path completo donde se encuentra el fichero XML. La mayoría de los campos estarán vacíos o a NULL, pero no es ningún inconveniente.

5.2. Creación de nuevas Tablas

Para no modificar las tablas cargadas del SIIU, se crean unas tablas nuevas. Los nombres elegidos han sido las siguientes:

- "alf09-10" para las tablas del curso 2009-2010 con el filtro I.
- "alf10-11" para las tablas del curso 2010-2011 con el filtro I.
- "j1_11-12" para las tablas del curso 2011-2012 con el filtro II.
- "j1_11-12" para las tablas del curso 2012-2013 con el filtro II.
- "j2_11-12" para las tablas del curso 2011-2012 con el filtro III.
- "j2_12-13" para las tablas del curso 2012-2013 con el filtro III.

Para incluir las nuevas tablas, basta con copiar la siguiente sentencia SQL:

```
CREATE TABLE `alf09-10` AS
SELECT `acc09-10`.*
FROM `acc09-10`, `rend09-10`
WHERE `acc09-10`.`NumeroDocumento`= `rend09-10`.`NumeroDocumento`
AND `rend09-10`.`Titulacion` = 2500397 AND
`rend09-10`.anioaccesosue = 2009;

CREATE TABLE `alf10-11` AS
SELECT `acc10-11`.*
FROM `acc10-11`, `rend10-11`
WHERE `acc10-11`.`NumeroDocumento`= `rend10-11`.`NumeroDocumento`
AND `rend10-11`.`Titulacion` = 2500397 AND
`rend10-11`.anioaccesosue = 2010;

CREATE TABLE `j2_11-12` AS
SELECT `acc11-12`.*
FROM `acc11-12`, `rend11-12`
WHERE `acc11-12`.`NumeroDocumento`= .`rend11-12`.`NumeroDocumento`
AND `rend11-12`.`Titulacion` = 2500397;

CREATE TABLE `j2_12-13` AS
SELECT `acc12-13`.*
FROM `acc12-13`, `rend12-13`
WHERE `acc12-13`.`NumeroDocumento`= .`rend12-13`.`NumeroDocumento`
AND `rend12-13`.`Titulacion` = 2500397;
```

```
CREATE TABLE `j1_11-12` AS
SELECT `acc11-12`.*
FROM `acc11-12`
WHERE `acc11-12`.`Titulacion` = 2500397;

CREATE TABLE `j1_12-13` AS
SELECT `acc12-13`.*
FROM `acc12-13`
WHERE `acc12-13`.`Titulacion` = 2500397;
```

5.3. Obtención de nuevas variables

Las nuevas variables obtenidas serán creadas dentro de las tablas recién creadas con las siguientes líneas SQL:

——————————REZAGADO——————————

```
alter table `alf09-10` add Rezagado int default 0;
alter table `alf10-11` add Rezagado int default 0;
alter table `j2_11-12` add Rezagado int default 0;
alter table `j2_12-13` add Rezagado int default 0;
alter table `j1_11-12` add Rezagado int default 0;
alter table `j1_12-13` add Rezagado int default 0;

Update `alf09-10` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `alf09-10` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;

Update `alf10-11` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `alf10-11` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;

Update `j2_11-12` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `j2_11-12` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;
```

```
Update `j2_12-13` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `j2_12-13` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;


Update `j1_11-12` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `j1_11-12` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;

Update `j1_12-13` set Rezagado = 1
where (AnioAccesoSUE - AnioNacimiento) = 19;
Update `j1_12-13` set Rezagado = 2
where (AnioAccesoSUE - AnioNacimiento) >=20;
```

CENTROSECMADRID

```
alter table `alf09-10` add CentroSecMadrid int default 0;
alter table `alf10-11` add CentroSecMadrid int default 0;
alter table `j2_11-12` add CentroSecMadrid int default 0;
alter table `j2_12-13` add CentroSecMadrid int default 0;
alter table `j1_11-12` add CentroSecMadrid int default 0;
alter table `j1_12-13` add CentroSecMadrid int default 0;

Update `alf09-10` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `alf09-10` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;


Update `alf10-11` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `alf10-11` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;
```

```
Update `j2_11-12` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `j2_11-12` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;


Update `j2_12-13` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `j2_12-13` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;


Update `j1_11-12` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `j1_11-12` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;


Update `j1_12-13` set CentroSecMadrid = 1
where MunicipioCentroSec < 28000 or MunicipioCentroSec > 29000;

Update `j1_12-13` set CentroSecMadrid = NULL
where MunicipioCentroSec = 88888
or MunicipioCentroSec = 99999 or MunicipioCentroSec = NULL;
```

——————TRABAJAESTUDIANTE——————————

```
alter table `alf09-10` add TrabajaEstudiante int default 1;
alter table `alf10-11` add TrabajaEstudiante int default 1;
alter table `j2_11-12` add TrabajaEstudiante int default 1;
alter table `j2_12-13` add TrabajaEstudiante int default 1;
alter table `j1_11-12` add TrabajaEstudiante int default 1;
```

```
alter table `j1_12-13` add TrabajaEstudiante int default 1;

Update `alf09-10` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `alf09-10` set TrabajaEstudiante = NULL
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;


Update `alf10-11` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `alf10-11` set TrabajaEstudiante = NULL
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;


Update `j2_11-12` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `j2_11-12` set TrabajaEstudiante = NULL
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;


Update `j2_12-13` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `j2_12-13` set TrabajaEstudiante = NULL
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;


Update `j1_11-12` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `j1_11-12` set TrabajaEstudiante = NULL
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;


Update `j1_12-13` set TrabajaEstudiante = 0
where OcupacionEstudiante is NULL;

Update `j1_12-13` set TrabajaEstudiante = NULL
```

```
where OcupacionEstudiante = 88 or OcupacionEstudiante = 99;
```

SEXONUM

```
alter table `alf09-10` add SexoNum int default 0;
alter table `alf10-11` add SexoNum int default 0;
alter table `j2_11-12` add SexoNum int default 0;
alter table `j2_12-13` add SexoNum int default 0;
alter table `j1_11-12` add SexoNum int default 0;
alter table `j1_12-13` add SexoNum int default 0;

Update `alf09-10` set SexoNum = 1 where SEXO = 'M';
Update `alf10-11` set SexoNum = 1 where SEXO = 'M';
Update `j2_11-12` set SexoNum = 1 where SEXO = 'M';
Update `j2_12-13` set SexoNum = 1 where SEXO = 'M';
Update `j1_11-12` set SexoNum = 1 where SEXO = 'M';
Update `j1_12-13` set SexoNum = 1 where SEXO = 'M';
```

ESPANOL

```
alter table `alf09-10` add Espanol int default 1;
alter table `alf10-11` add Espanol int default 1;
alter table `j2_11-12` add Espanol int default 1;
alter table `j2_12-13` add Espanol int default 1;
alter table `j1_11-12` add Espanol int default 1;
alter table `j1_12-13` add Espanol int default 1;


Update `alf09-10` set Espanol = 0 where `PaisNacionalidad` = 724;
Update `alf10-11` set Espanol = 0 where `PaisNacionalidad` = 724;
Update `j2_11-12` set Espanol = 0 where `PaisNacionalidad` = 724;
Update `j2_12-13` set Espanol = 0 where `PaisNacionalidad` = 724;
Update `j1_11-12` set Espanol = 0 where `PaisNacionalidad` = 724;
Update `j1_12-13` set Espanol = 0 where `PaisNacionalidad` = 724;
```

EDAD

```
alter table `alf09-10` add Edad int;
alter table `alf10-11` add Edad int;
alter table `j2_11-12` add Edad int;
alter table `j2_12-13` add Edad int;
alter table `j1_11-12` add Edad int;
alter table `j1_12-13` add Edad int;

Update `alf09-10` set Edad = '2009' - `AnioNacimiento`;
Update `alf10-11` set Edad = '2010' - `AnioNacimiento`;
Update `j2_11-12` set Edad = '2011' - `AnioNacimiento`;
Update `j2_12-13` set Edad = '2012' - `AnioNacimiento`;
Update `j1_11-12` set Edad = '2011' - `AnioNacimiento`;
Update `j1_12-13` set Edad = '2012' - `AnioNacimiento`;
```

RANGO NOTA

```
alter table `alf09-10` add RangoNota int;
alter table `alf10-11` add RangoNota int;
alter table `j2_11-12` add RangoNota int;
alter table `j2_12-13` add RangoNota int;
alter table `j1_11-12` add RangoNota int;
alter table `j1_12-13` add RangoNota int;

Update `alf09-10` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6;
Update `alf10-11` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6;
Update `j2_11-12` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6 ;
Update `j2_12-13` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6 ;
Update `j1_11-12` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6 ;
Update `j1_12-13` set RangoNota = 0
where NotaAdmision >= 5 and NotaAdmision < 6 ;

Update `alf09-10` set RangoNota = 1
where NotaAdmision >= 6 and NotaAdmision < 7;
Update `alf10-11` set RangoNota = 1
where NotaAdmision >= 6 and NotaAdmision < 7;
Update `j2_11-12` set RangoNota = 1
```

```
where NotaAdmision >= 6 and NotaAdmision < 7;
Update `j2_12-13` set RangoNota = 1
where NotaAdmision >= 6 and NotaAdmision < 7 ;
Update `j1_11-12` set RangoNota = 1
where NotaAdmision >= 6 and NotaAdmision < 7;
Update `j1_12-13` set RangoNota = 1
where NotaAdmision >= 6 and NotaAdmision < 7 ;

Update `alf09-10` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8;
Update `alf10-11` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8;
Update `j2_11-12` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8;
Update `j2_12-13` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8 ;
Update `j1_11-12` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8;
Update `j1_12-13` set RangoNota = 2
where NotaAdmision >= 7 and NotaAdmision < 8 ;

Update `alf09-10` set RangoNota = 3 where NotaAdmision >= 8    ;
Update `alf10-11` set RangoNota = 3 where NotaAdmision >= 8   ;
Update `j2_11-12` set RangoNota = 3 where NotaAdmision >= 8   ;
Update `j2_12-13` set RangoNota = 3 where NotaAdmision >= 8   ;
Update `j1_11-12` set RangoNota = 3 where NotaAdmision >= 8   ;
Update `j1_12-13` set RangoNota = 3 where NotaAdmision >= 8   ;

Update `alf09-10` set RangoNota = NULL where NotaAdmision IS NULL
;
Update `alf10-11` set RangoNota = NULL where NotaAdmision IS NULL
;
Update `j2_11-12` set RangoNota = NULL where NotaAdmision IS NULL
;
Update `j2_12-13` set RangoNota = NULL where NotaAdmision IS NULL
;
Update `j1_11-12` set RangoNota = NULL where NotaAdmision IS NULL
;
Update `j1_12-13` set RangoNota = NULL where NotaAdmision IS NULL
;
```

5.4. Correcciones

Las variables obtenidas tienen valores de tipo "no consta" o "no aplica". En este proyecto se ha decidido tratar estos valores como "NULL", para así hacerlos más uniformes. Las siguientes líneas SQL corrigen dicho problema:

```
UPDATE `alf09-10` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;
UPDATE `alf10-11` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;
UPDATE `j2_11-12` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;
UPDATE `j2_12-13` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;
UPDATE `j1_11-12` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;
UPDATE `j1_12-13` SET `NotaAdmision`= NULL WHERE NotaAdmision > 88;


UPDATE `alf09-10` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';
UPDATE `alf10-11` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';
UPDATE `j2_11-12` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';
UPDATE `j2_12-13` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';
UPDATE `j1_11-12` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';
UPDATE `j1_12-13` SET `NivelEstudioPadre` = NULL
WHERE `NivelEstudioPadre` = '8' OR `NivelEstudioPadre` = '9';

UPDATE `alf09-10` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';
UPDATE `alf10-11` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';
UPDATE `j2_11-12` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';
UPDATE `j2_12-13` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';
UPDATE `j1_11-12` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';
UPDATE `j1_12-13` SET `NivelEstudioMadre` = NULL
WHERE `NivelEstudioMadre` = '8' OR `NivelEstudioMadre` = '9';

UPDATE `alf09-10` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';
```

```
UPDATE `alf10-11` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';
UPDATE `j2_11-12` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';
UPDATE `j2_12-13` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';
UPDATE `j1_11-12` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';
UPDATE `j1_12-13` SET `OcupacionPadre`= NULL
WHERE `OcupacionPadre`= '88' OR `OcupacionPadre`= '99';

UPDATE `alf09-10` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';
UPDATE `alf10-11` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';
UPDATE `j2_11-12` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';
UPDATE `j2_12-13` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';
UPDATE `j1_11-12` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';
UPDATE `j1_12-13` SET `OcupacionMadre`= NULL
WHERE `OcupacionMadre`= '88' OR `OcupacionMadre`= '99';

UPDATE `alf09-10` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';
UPDATE `alf10-11` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';
UPDATE `j2_11-12` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';
UPDATE `j2_12-13` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';
UPDATE `j1_11-12` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';
UPDATE `j1_12-13` SET `OcupacionEstudiante`= NULL
WHERE `OcupacionEstudiante`= '88' OR `OcupacionEstudiante`= '99';

UPDATE `alf09-10` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';
UPDATE `alf10-11` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';
```

```
UPDATE `j2_11-12` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';
UPDATE `j2_12-13` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';
UPDATE `j1_11-12` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';
UPDATE `j1_12-13` SET `EspecialidadAcceso`= NULL
WHERE `EspecialidadAcceso`= '8888';

UPDATE `alf09-10` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';
UPDATE `alf10-11` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';
UPDATE `j2_11-12` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';
UPDATE `j2_12-13` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';
UPDATE `j1_11-12` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';
UPDATE `j1_12-13` SET `PaisFinEstudioAcceso`= NULL
WHERE `PaisFinEstudioAcceso`= '999';

UPDATE `alf09-10` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';
UPDATE `alf10-11` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';
UPDATE `j2_11-12` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';
UPDATE `j2_12-13` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';
UPDATE `j1_11-12` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';
UPDATE `j1_12-13` SET `EstudioAcceso`= NULL
WHERE `EstudioAcceso`= '88' OR `EstudioAcceso`= '99';

UPDATE `alf09-10` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
UPDATE `alf10-11` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
UPDATE `j2_11-12` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
```

```
UPDATE `j2_12-13` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
UPDATE `j1_11-12` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
UPDATE `j1_12-13` SET `NaturalezaCentroSec`= NULL
WHERE`NaturalezaCentroSec`= '9';
```

5.5. Variable Abandono

La variable Abandono también debe ser añadida. Dada la complejidad de las sentencias SQL necesarias para obtenerlas, se ha dedicado una sección exclusiva para ello. Las líneas SQL necesarias para añadir la variable abandono son las siguientes:

```
ALTER TABLE `alf09-10` ADD abandono int default 0;
ALTER TABLE `alf10-11` ADD abandono int default 0;
ALTER TABLE `j2_11-12` ADD abandono int default 0;
ALTER TABLE `j2_12-13` ADD abandono int default 0;
ALTER TABLE `j1_11-12` ADD abandono int default 0;
ALTER TABLE `j1_12-13` ADD abandono int default 0;
```

```
Update `alf09-10` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc09-10`.NumeroDocumento
FROM `acc09-10`, `rend09-10`
WHERE `acc09-10`.`NumeroDocumento`= .`rend09-10`.`NumeroDocumento`
AND `rend09-10`.`Titulacion` = 2500397
AND `rend09-10`.anioaccesosue = 2009
AND (((`acc09-10`.NumeroDocumento)
   NOT IN (SELECT `rend10-11`.NumeroDocumento
FROM `rend10-11` )))

);

Update `alf10-11` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc10-11`.NumeroDocumento
FROM `acc10-11`, `rend10-11`
```

```
WHERE  `acc10-11`.`NumeroDocumento`= .`rend10-11`.`NumeroDocumento`
AND `rend10-11`.`Titulacion` = 2500397
AND `rend10-11`.anioaccesosue = 2010
AND (((`acc10-11`.NumeroDocumento)
   NOT IN (SELECT `rend11-12`.NumeroDocumento
FROM `rend11-12`  )))


);
Update `j2_11-12` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc11-12`.NumeroDocumento
FROM `acc11-12`,  `rend11-12`
WHERE  `acc11-12`.`NumeroDocumento`= .`rend11-12`.`NumeroDocumento`
AND `rend11-12`.`Titulacion` = 2500397
AND (((`acc11-12`.NumeroDocumento)
   NOT IN (SELECT `rend12-13`.NumeroDocumento
FROM `rend12-13`  )))

);


Update `j2_12-13` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc12-13`.NumeroDocumento
FROM `acc12-13`,  `rend12-13`
WHERE  `acc12-13`.`NumeroDocumento`= .`rend12-13`.`NumeroDocumento`
AND `rend12-13`.`Titulacion` = 2500397
AND (((`acc12-13`.NumeroDocumento)
   NOT IN (SELECT `ava13-14`.NumeroDocumento
FROM `ava13-14`  )))

);


Update `j1_11-12` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc11-12`.NumeroDocumento
FROM `acc11-12`
WHERE `acc11-12`.titulacion = 2500397
AND(((`acc11-12`.NumeroDocumento)
```

```
    NOT IN (SELECT `rend12-13`.NumeroDocumento
FROM `rend12-13`  )))

);


Update `j1_12-13` set abandono = 1 where `NumeroDocumento` IN (

SELECT `acc12-13`.NumeroDocumento
FROM `acc12-13`
WHERE `acc12-13`.titulacion = 2500397
AND(((`acc12-13`.NumeroDocumento)
    NOT IN (SELECT `ava13-14`.NumeroDocumento
FROM `ava13-14`  )))

);
```

5.6. Unión de Tablas

Por último, es necesario unir los distintos cursos en una nueva tabla. De esta manera se obtendrán dos nuevas tablas:

- La tabla "alfj1_09–13", que contendrá los ficheros de los cursos 2009-2010, 2010-2011 (ambos con el filtro de tipo I), 2011-2012 y 2012-2013 (ambos con el filtro de tipo II).
- La tabla "alfj2_09–13", que contendrá los ficheros de los cursos 2009-2010, 2010-2011 (ambos con el filtro de tipo I), 2011-2012 y 2012-2013 (ambos con el filtro de tipo III)

```
create table `alfj2_09--13` select *
from(
SELECT * FROM `alf09-10`
UNION
SELECT * FROM `alf10-11`
UNION
SELECT * FROM `j2_11-12`
UNION
SELECT * FROM `j2_12-13`
) a;
```

```
create table `alfj1_09--13` select *
from(
SELECT * FROM `alf09-10`
UNION
SELECT * FROM `alf10-11`
UNION
SELECT * FROM `j1_11-12`
UNION
SELECT * FROM `j1_12-13`
) a
```

5.7. Modelado

Dada la programación modular y visual utilizada para programar en Knime y Clementine, no ha sido posible una sistematización de esta parte. Sería posible transcribir a Java el modelado realizado con Knime, pero debido a la complejidad de dicho problema, quedará como una línea futura. En la figura 5.1 se puede observar una parte del área de trabajo de Knime. Además, la dificultad de la sistematización del modelo aumenta debido a que cada año se introducirán nuevos datos y esto puede suponer no solo nuevas decisiones en el modelo sino que también nuevos mecanismos de preprocesamiento.

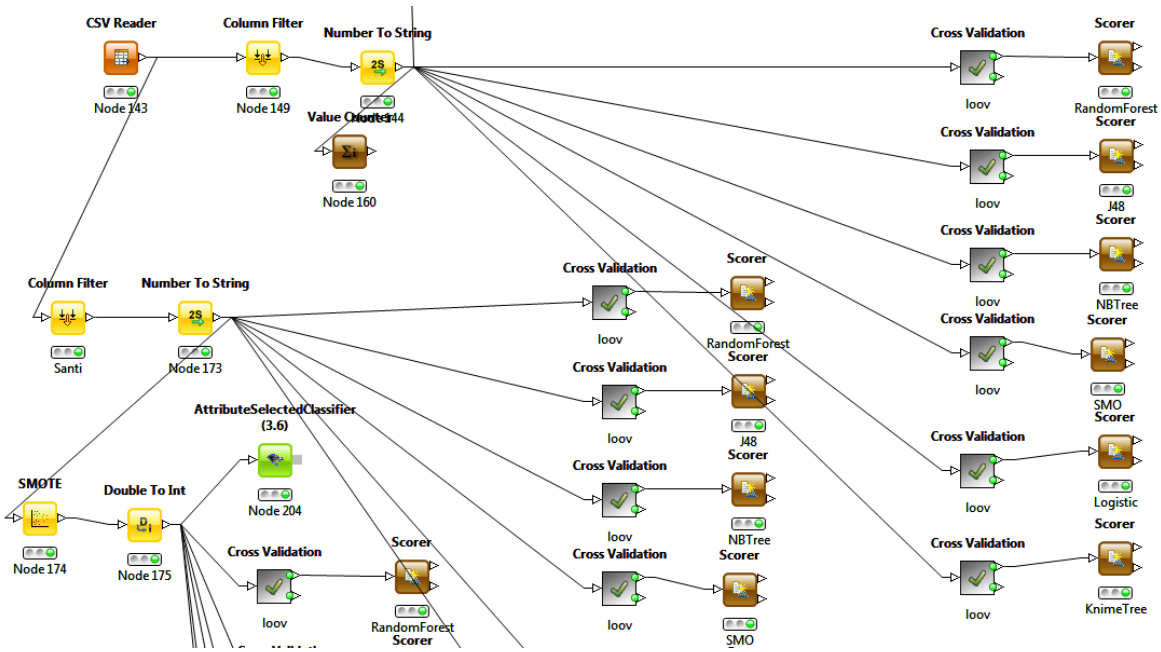

Figura 5.1: Muestra del área de trabajo de Knime

Capítulo 6

Conclusiones y líneas futuras

Los modelos hallados son buenos desde el punto de vista informático, ya que tienen un porcentaje de acierto en torno al 80 %, minimizando aquellos alumnos clasificados como *no abandono* que finalmente abandonan. Es de suponer que con el paso de los años y, por lo tanto, teniendo más cantidad de datos, estos modelos serán más fiables.

6.1. Dificultades encontradas

Ha habido distintas dificultades a lo largo de todo el proyecto, como en la obtención de los datos del SIIU o en el modelado, pero, sin lugar a dudas, la mayor dificultad se encuentra en la definición de abandono. Han sido necesarios varios meses para llegar a un acuerdo en dicha definición. El problema radica en la existencia de tres ficheros distintos (Acceso, Avance y Rendimiento) de donde se obtenían los alumnos, dado que las diferentes definiciones de abandono dependían de qué fichero se analizaba primero o del orden de los mismos. Una vez resuelto este problema, el resto de dificultades no han sido comparables.

6.2. Conocimientos adquiridos

La realización de este proyecto ha sido una buena conclusión del Grado de Ingeniería Informática, dado que ha tocado aquellos campos que no se han visto en el grado. En los cuatro años que corresponden a dicho grado se han aprendido a desenvolverse en muchos aspectos, sin miedo a abordar cualquier tipo de problema. El desconocimiento sobre el tema principal del proyecto al comienzo del mismo, *Data Mining*, no ha sido un impedimento sino un aliciente para llevarlo a cabo. Los conocimiento adquiridos a lo largo del grado, dado que son más amplios que profundos, han sido un instrumento óptimo para enfrentarse a cualquier tipo de problema.

Más concretamente, los conocimientos adquiridos por la realización del proyecto son los siguientes:

- Se ha aprendido a realizar un proyecto de *Data Mining*, llevando a cabo la metodología CRISP-DM.
- La mayoría de algoritmos utilizados para el modelo eran completamente desconocidos. Con la realización del proyecto, el conocimiento sobre los mismos ha aumentado notablemente, adquiriendo la capacidad de analizar por qué un algoritmo es mejor que otro en algunos aspectos.
- El uso de herramientas de bases de datos era más teórico que práctico, así como del lenguaje SQL. Con este proyecto, dichos conocimientos son más amplios y profundos.
- El conocimiento de programación modular era únicamente teórico y aunque la programación en Clementine y Knime es modular, no ha habido problema alguno a la hora de comprender el flujo o el significado de los módulos o nodos utilizados.
- Por último, cabe destacar el conocimiento nulo del sistema de composición de textos LaTeX a la hora de comenzar el proyecto, pero que a la finalización del mismo, se puede defender a la hora de realizar documentos con dicha herramienta.

6.3. Posibles utilidades

Uno de los puntos que más motivación ha otorgado ha sido que el proyecto no es sólo un proyecto teórico, sino que es un punto de partida que tiene como fin la reducción del abandono en la Escuela Técnica Superior de Ingenieros Informáticos. La posibilidad de trabajar directamente con el director de la escuela Víctor Robles, ha sido no sólo un orgullo sino que también un aliciente, puesto que sus distintas explicaciones y ambiciones a la hora de reducir el abandono a través de la motivación a los estudiantes han hecho que el esfuerzo a la hora de realizar el proyecto sea máximo. En definitiva, el fin de este proyecto es detectar aquellos alumnos que tienen mayor posibilidad de abandonar para así realizar políticas que puedan motivarlos.

6.4. Experiencia personal

La experiencia ha sido satisfactoria y muy grata. Si bien es cierto que se ha echado de menos algo de implementación, esto no ha sido un obstáculo motivacional, simplemente un punto de vista distinto a la hora de contemplar un problema de un ingeniero informático: no todo es programar. Esta experiencia es únicamente el principio

de una carrera profesional y todo el conocimiento adquirido no sólo estos meses, sino que también estos años de grado será de utilidad a la hora de madurar personal y profesionalmente.

6.5. Líneas futuras

Existen varias líneas abiertas para futuras investigaciones que deben de ser tratadas en los meses próximos:

- La técnica utilizada para igualar el número de alumnos que abandonan con aquellos que no abandonan ha sido "oversampling", que consiste en ampliar artificialmente el número de instancias que abandonan. Una línea futura sería en utilizar otros métodos alternativos para igualar el número de instancias que abandonan con aquellas que no abandonan, como el "sampling", que consiste en eliminar instancias de la variable predominante.

- Dado que únicamente se dispone de cuatro cursos en el conjunto de datos, sería muy interesante que el modelo aprendiese de tres de ellos y validase con el cuarto, para para así comprobar la eficacia del mismo. Además, cuantos más años se utilicen para realizar el modelo, mejor va a ser éste, por lo que sería conveniente realizar el modelo en cursos futuros.

- Aunque el modelo ha sido probado exclusivamente con los alumnos de la Escuela Técnica Superior de Ingenierieros Informáticos, también sería posible el uso en el resto de escuelas de la Universidad Politécnica de Madrid, por lo que otra línea futura es ampliar el proyecto al resto de escuelas y facultades.

- A pesar de la complejidad del proyecto y aunque se ha intentado sistematizar, el hecho de haber encontrado un modelo capaz de predecir el abandono con semejante porcentaje anima a la automatización del proyecto, desarrollando en Java la programación modular realizada con Knime y mediante *scripts* la parte relativa al manejo de la base de datos.

Capítulo 7

Referencias

[**García13**] A. García, J. Blanco, A. Casaravilla, A. Castejón, A. Gonzalo, M. A. Mahíllo, M. Malinga *Protocolo de Calidad para la Tasa de Permanencia a un Año en la UPM* III CLABES pg 985-995 ISBN: 978-84-15302-71-1 2013

[**Frawley96**] W. J. Frawley, G. Piatestky-Shapiro, P. Smyth *From Data Mining to Kowledge Discovery: An Overview Advances in Knowledge Discovery and Data Mining.* ed. W. J. Frawley, G. Piatestky-Shapiro, P. Smyth and Uthurusamy, AAAI Press 1996

[**Piatesky91**] G. Piatesky-Shaphiro, *Report on the AAAI-91 Workshop on Knowledge Discovery in Databases*, IEEE Expert 6(5), October, pp. 74-76.

[**Fayyad 96**] U. Fayyad, G. Piatestky-Shaphiro, P. Smyth *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, In Proceedings The Second International Conference on Knowledge discovery and Data Mining, August 1996, pp. 82-88.

[**Cabena 97**] P. Cabena *Discovering Data Mining: From Concept to Implementation*, IBM, 1997, ISBN 01-374-3980-6.

[**Wirth00**] R. Wirth, J. Hipp *CRISP-DM: Towards a Standard Process Model for Data Mining*, 2000

[**Kohavi95**] R. Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, In International Joint Conference on Artificial Intelligence (IJCAI), 1995

[**Rahman13**] M. Mostafizur Rahman and D. N. Davis, *Addressing the Class Imbalance Problem in Medical Datasets*, In International Journal of Machine Learning and Computing, Vol. 3, No. 2, April 2013

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
	Fecha/Hora	Fri Jun 06 20:37:53 CEST 2014
	Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
	Numero de Serie	630
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)